Univerzita Karlova

Filozofická fakulta

Ústav informačních studií a knihovnictví

Diplomová práce

Bc. Marie Kazárová

Získávání znalostí z marketingových dat

Knowledge discovery in marketing data

Praha 2019

Vedoucí práce: prof. RNDr. Jiří Ivánek, CSc.

Poděkování:
Velice ráda bych v prvé řadě poděkovala svému vedoucímu diplomové práce prof. RNDr. Jiřímu Ivánkovi, CSc., za odborné vedení této práce, cenné rady, trpělivost, ochotu a čas, který mi při vedení práce věnoval. Také bych chtěla poděkovat svému muži a rodině za jejich podporu po celou dobu, co práce vznikala. Poděkovat bych chtěla i Petře Černohlávkové za její důkladnou zpětnou vazbu.

Prohlášení:

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, dne 16. 12. 2019

Bc. Marie Kazárová

Klíčová slova:

dobývání znalostí z databází; data mining; CRISP-DM; marketing; online marketing

Keywords:

knowledge discovery in databases; data mining; CRISP-DM; marketing; online marketing

Abstrakt

Data miningové techniky jsou v komerční sféře využívány za účelem získávání konkurenčních výhod. V oblasti marketingu v současnosti především v souvislosti s personalizací reklamy a udržení dlouhodobých vztahů se zákazníky. Vývoj v oblasti dobývání znalostí z databází, v kombinaci s trvalým růstem výpočetního výkonu a jeho dostupností, přináší nejen pozitivní dopady, ale i značná úskalí. V praktické aplikaci diplomové práce jsou s využitím data miningových technik ověřeny možnosti získávání znalostí z dat webové analytiky v kombinaci s transakčními daty e-commerce společností. Cílem experimentální aplikace je ověřit, zda existuje segment uživatelů, kteří reagují na marketingovou komunikaci častěji, než jiné segmenty, a nalézt případné souvislosti v databázi. Pomocí data miningové techniky shlukování je takový segment v datech nalezen. Součástí praktické aplikace je i klasifikační model na bázi rozhodovacích stromů, který s přesností 75% určí, zda uživatel provede transakci, či ne. Tento typ výstupu lze následně využít k optimalizaci marketingové a obchodní strategie.

Abstract

Data mining techniques are used by companies to gain competitive advantages. In today's marketplace, they are also used by marketers mainly for personalization of advertising and for maintaining long-term relationship with customers. Progress in knowledge discovery in databases and availability of computational power comes not only with positive impact, but also with challenges. The practical part of the thesis aims to explore and describe data mining techniques applied to e-commerce dataset. Dataset consists of transaction and web analytics data. The goal of experimental application aims to make a selection of users who most probably react to a marketing communication and to identify the factors which influence them. Target segment of users is obtained through the use of data mining technique clustering. The classification model uses decision tree algorithm to predict whether users submit transaction with an accuracy of 75%. The results are useful for optimization of marketing and business strategy.

OBSAH

Předmluva

Výběr tématu diplomové práce souvisí s mojí delší pracovní zkušeností na pozici datového analytika v marketingové oblasti, kdy jsem při psaní diplomové práce čerpala mimo jiné i z těchto zkušeností. Název diplomové práce, *Získávání znalostí z dat*, je obsahovým ekvivalentem k označení *Dobývání znalostí z databází (Knowledge discovery in databases, KDD)*. Přestože je problematika dobývání znalostí z databází nezávislá na odvětví původu dat, z důvodu mé přímé zkušenosti s marketingovými daty jsem se ve své diplomové práci zaměřila na oblast marketingu a e-commerce.

Během svých pracovních zkušeností jsem narazila na možnosti i limity dobývání znalostí z databází a zjistila, jak moc je tento proces komplexní a s jakými problémy se lze setkat. V teoretické části diplomové práce jsem těchto zkušeností využila při uvedení problematiky do kontextu s praxí, při analýze metodologií, které pro dobývání znalostí z databází od 90.let minulého století v tomto oboru vznikly a pro výběr vhodných technik a nástrojů pro *data miningové* (*DM*) úlohy v kontextu marketingu. V diplomové práci jsem kompilovala dostupné informace, odborné knihy a odborné články se svými vlastními zkušenostmi v oblasti získávání znalostí z marketingových dat. Součástí práce jsou i vybrané případové studie využití *data miningových* technik v oblasti marketingu.

V rámci experimentální aplikace jsem dle postupu metodologie *Cross-industry standard process for data mining* (*CRISP-DM)* řešila hypotézu, zda existuje skupina uživatelů internetu, kteří na marketingovou komunikaci reagují častěji, konkrétně uskutečněním transakce na Google Merchandise Store, než jiné skupiny, s nalezením případných souvislostí v databázi za účelem optimalizace marketingové a obchodní strategie. Součástí výstupů praktické aplikace jsou doporučení, jak lze získané znalosti dále využít. Ke splnění řešení je využit programovací jazyk R, platforma pro data mining RapidMiner, vizualizační platforma Tableau a cloudová platforma Google BigQuery s volně dostupným datasetem Google Analytics Sample obsahující data webové analytiky z Google Analytics 360 a transakční data z Google Merchandise Store. Jedná se o vzorek dat typický pro oblast e-commerce, který zahrnuje informace, odkud zákazník na web přišel a popisuje jeho cestu a interakci na webu až do fáze případného dokončení nákupu. Vzhledem k tomu, že se jedná o vzorová data Google Analytics, která zahrnují transakce z několika e-shopů z

celého světa za období jednoho roku, je nutné brát výsledky praktické části pouze jen jako orientační a slouží spíše k demonstraci možností praktického využití *data miningových* technik v oblasti marketingu. Nelze z nich odvodit obecné zákonitosti.

Vzhledem k anglicko-české nejednoznačnosti v terminologii jsem se pro označování používané v diplomové práci inspirovala u P. Berky (Berka 2003, s.12). Termín *Dobývání znalostí z databází* je ekvivalentem pro Knowledge discovery in databases (*KDD*). Termín *data mining* (ve shodě s metodologií *CRISP-DM*) je ekvivalentem pro modelování, analytické metody nebo analytické procedury.

Citace v diplomové práci jsou dle normy ČSN ISO 690.

Rozsah textu diplomové práce je 135 049 znaků (75NS).

1. Úvod

Od počátků oboru *Dobývání znalostí z databází (KDD)* za poslední tři dekády došlo k jeho rapidnímu vývoji a poptávka po znalostech a zkušenostech v oboru exponenciálně roste. V 90. letech došlo k základnímu vymezení procesu, pojmů a terminologie. Obor se neustále vyvíjí na poli komerční i vědecké oblasti. Z důvodu rychlosti vývoje a široké oblasti dalších oborů, kterých se *KDD* dotýká, se významy a terminologie pojmů mohou v různých publikacích lišit. Vznik velkého množství dalších nových pojmů s nesjednocenou terminologií a informační přehlcenost může být úskalím oboru. Zmatení pojmů a popularizace oboru může přinést v důsledku nereálná očekávání zadavatelů již ve fázi vstupních požadavků. Z tohoto důvodu je důležité vymezení pojmů *Dobývání znalostí z databází* a *Data mining,* jejich definic a vztahu pojmu *Data mining* k procesu *KDD*.

KDD procesy jsou od jeho počátků uplatněny v komerční sféře za účelem získávání konkurenčních výhod. V oblasti marketingu jsou využívány pro lepší interakci se zákazníkem a pro tvorbu obchodních a kampaňových strategií. V současnosti je v oblasti marketingu nástrojem pro de-masifikaci kampaňových strategií či pro de-masifikaci komunikace se zákazníky. Tyto obchodní a marketingové strategie jsou pro společnosti mnohem ziskovější, než plošné cílení reklamy na velké skupiny lidí. Za účelem metodologické podpory procesu vznikly jak metodologie od producentů, tak i univerzální metodologie *CRISP-DM*, které ve svých standardech sdílí dosavadní zkušenosti a poznatky v oboru.

Stejně tak, jako je v procesu klíčová správná definice problému na vstupu, čili správná definice řešené otázky, je neméně klíčová volba vhodné modelovací techniky pro získání relevantních znalostí z dat na výstupu. Ne všechny modelovací techniky jsou vhodné pro různé typy oblastí, ze kterých data pochází. V oblasti marketingu jsou často využity techniky rozhodovacích stromů za účelem klasifikace a techniky shlukování za účelem segmentace uživatelů. Cílem je z obchodního hlediska ve většině případů personalizace reklamy a udržení dlouhodobých vztahů se zákazníky, kterých, vzhledem k aktuálně velkému množství generovaných dat, nelze dosáhnout jinak, než s využitím informačních systémů a technik *data miningu*.

Hlavním cílem praktické aplikace diplomové práce je z obchodního hlediska nalézt, pokud existuje, skupinu uživatelů, kteří reagují na marketingové kampaně častěji, než ostatní, nalézt případné souvislosti v databázi, které mají na tuto skutečnost vliv, a tomu přizpůsobit marketingovou a obchodní strategii za účelem zvýšení konverzního poměru. K řešení je zvolena metoda segmentace s využitím techniky shlukování K-středů. Vedlejším obchodním cílem je odhadnout, zda nově akvirovaný uživatel provede transakci, či ne, a tomu již od počátku přizpůsobit komunikaci. Pro vytvoření odhadu, zda nový návštěvník provede transakci, či ne, je využita metoda klasifikace pomocí techniky rozhodovacích stromů.

2. Data mining

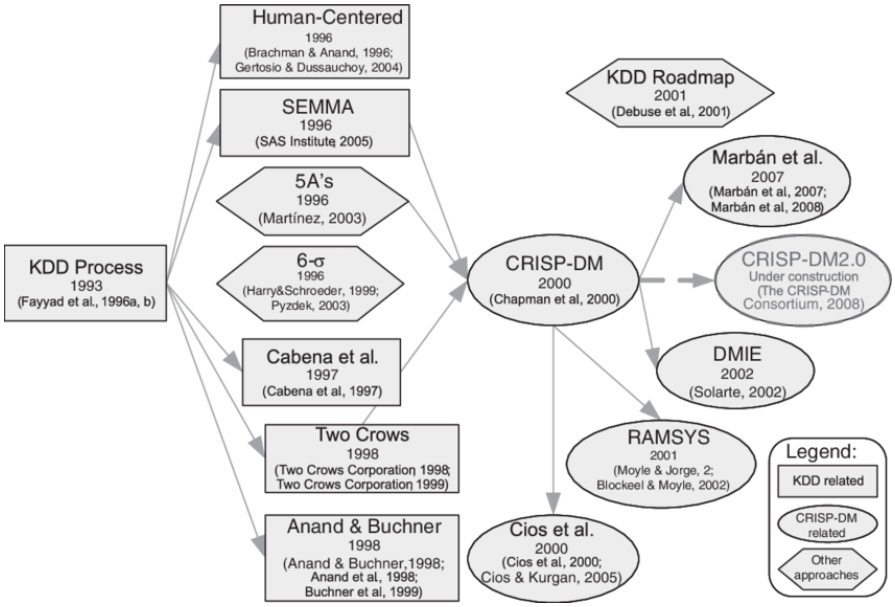
Techniky *data miningu* jsou součástí procesu *Dobývání znalostí z databází (KDD).* Kapitola obsahuje vymezení pojmů *Dobývání znalostí z databází* a *Data mining,* jejich definic a vztahu pojmu *Data mining* k procesu *KDD* a uvedení *KDD* procesu do kontextu praxe a oblasti marketingu. V kapitole jsou následně popsány metodologie procesu a to jak od producentů, tak univerzální metodologie *CRISP-DM.* Ne všechny data miningové techniky jsou vhodné pro různé oblasti a typy úloh. V kapitole jsou uvedeny a rozepsány vybrané *DM* techniky vhodné k využití v oblasti marketingu včetně uvedení těchto technik do kontextu oboru.

2.1. Dobývání znalostí z databází (KDD)

Od počátků oboru v 90.letech vzniklo velké množství nových pojmů s nesjednocenou terminologií. Z tohoto důvodu je důležité v rámci diplomové práce vymezit pojmy *Dobývání znalostí z databází* a *Data mining,* popsat jejich definice a vztahu pojmu *Data mining* k procesu *KDD*. Neustálý růst výpočetního výkonu a dostupnost informací v dnešní době má v kontextu oboru *KDD* nejen pozitivní, ale i negativní důsledky. *KDD* proces je za účelem konkurenční výhody v komerční sféře využíván již od jeho počátků a v souvislosti s marketingem je aktuálně využíván především pro personalizaci reklamy a udržení si stávajících zákazníků.

2.1.1. Definice KDD a Data miningu

Pojem *Dobývání znalostí z databází* lze definovat jako: "*Netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat"* (Berka, 2003, s.15). Fayyad a kolektiv v roce 1996 v této definici stanovuje (Fayyad et al., 1996, s.83), že proces *získávání* sestává ze spousty kroků, které se mnohokrát opakují. *Netriviální* znamená, že se nejedná o běžné výpočty jako je např. pouhé určení průměrné hodnoty. Objevené *informace* by měly být nové, dříve neznámé, na první pohled nepředvídatelné, zajímavé, srozumitelné, měly by být užitečné a aplikace na nových datech by měla přinášet s určitou mírou pravděpodobnosti validní výsledky. Z Fayyadova vymezení procesu *KDD* a terminologie následně vycházel další vývoj oboru, tuto skutečnost zrcadlil i vývoj metodologií *KDD* (viz Obrázek 1: Vývoj metodologií). Již v devadesátých letech byl narůst množství dat natolik značný, že vznikala urgentní poptávka po stále větším výpočetním výkonu a nástrojích pro získávání užitečných informací a následně i znalostí z dat. (Fayyad et al. 1996, s.82)



Obrázek 1: Vývoj metodologií. Dostupné online z: https://www.researchgate.net/figure/Evolution-of-data-mining-process-models-and-methodologies_fig4_220254274

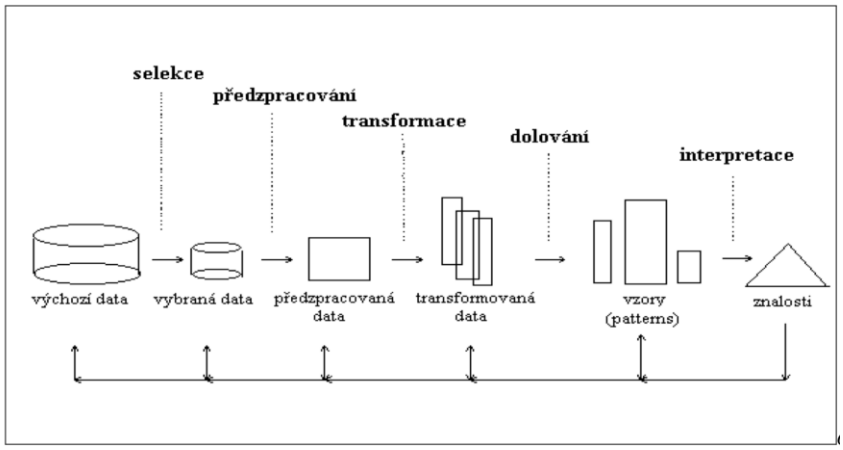
Termín *Dobývání znalostí z databází*, jehož anglickým ekvivalentem je *Knowledge Discovery in Databases (KDD)*, byl poprvé použit v roce 1989 během prvního mezinárodního workshopu (IJCAI-89) na téma *Dobývání znalostí z databází*. Pojem *KDD*

měl zdůrazňovat, že výstupem procesu objevů založených na datech (*data-driven discovery*) je znalost. Tento termín byl dále šířen i v oblastech umělé inteligence a strojového učení. *KDD* se vyvíjelo průnikem oblastí strojového učení, rozpoznávání vzorů, databází, statistiky, umělé inteligence, expertních systémů, datových vizualizací, s využitím tehdy dostupného výpočetního výkonu. Společným jmenovatelem bylo vždy získat znalost vysoké úrovně z dat nízké úrovně v rozsáhlých datových setech (Fayyad et al., 1996, s.82).

Historicky existovala různá označení, která zahrnovala celý proces hledání užitečných *vzorů* (patterns) v datech, včetně označení *dolování z dat* (data mining), *extrakce znalostí* (knowledge extraction), *objevování informací* (information discovery), *sklízení informací* (information harvesting), *datová archeologie* (data archeology) či *proces hledání vzorů v datech* (data pattern processing). Fayyad a kolektiv, s cílem sjednocení procesu a používané terminologie v roce 1996, zavádí označení *data mining* pouze pro část procesu *Dobývání znalostí z databází* s tím, že prostá aplikace specifického algoritmu pro nalezení informací v datech, bez fáze přípravy dat a interpretace, je zavádějící (Fayyad et al., 1996, s.82). Příprava dat a interpretace výsledných znalostí je tak přidanou hodnotou oproti samostatně užitým statistickým metodám a metodám strojového učení. Fayyad odlišuje *znalosti* (vzory) získané dolováním z dat od *znalostí* získaných interpretací uživatelem. Toto rozlišení znalostí, ve shodě s P. Berkou (Berka, 2003, s.16), v této diplomové práci prováděno nebude. *Znalostí* je zde shodně míněn výstup interpretace uživatelem.

Proces *KDD*, který v roce 1996 popsal Fayyad a kolektiv, je interaktivní proces zahrnující spoustu kroků a to přípravu dat, jejich výběr, předzpracování a transformaci, hledání vzorů a informací aplikováním data miningových metod, interpretaci těchto výstupů, získání znalostí, vylepšení a opětovné spuštění procesu. Proces probíhá v rámci několika iterací s využitím znalostí získaných v předchozích iteracích. Během procesu je po uživateli vyžadováno, aby učinil spoustu rozhodnutí (Fayyad et al., 1996, s.83). Celkový proces je znázorněn na Obrázku 2 (Proces KDD dle Fayyad et al., 1996). Tento proces v té době nebyl jediným. Podobné návrhy již byly předloženy v oblasti statistiky (Hand, 1994 podle Fayyad et al., 1996, s.83) a v oblasti strojového učení (Brodley, Smith, 1996 podle Fayyad et al., 1996, s.83). V současné době je možné dohledat různé modifikace tohoto postupu, ze kterých postupně vznikaly různé metodologie, viz Obrázek 1 (Vývoj metodologií), avšak vždy byla zachována logická struktura postupu, kterou zavedl Fayyad a kolektiv. V jednom z nich byl například zmenšen počet fází na tři, zkráceně na *preprocessing*, *data mining* a

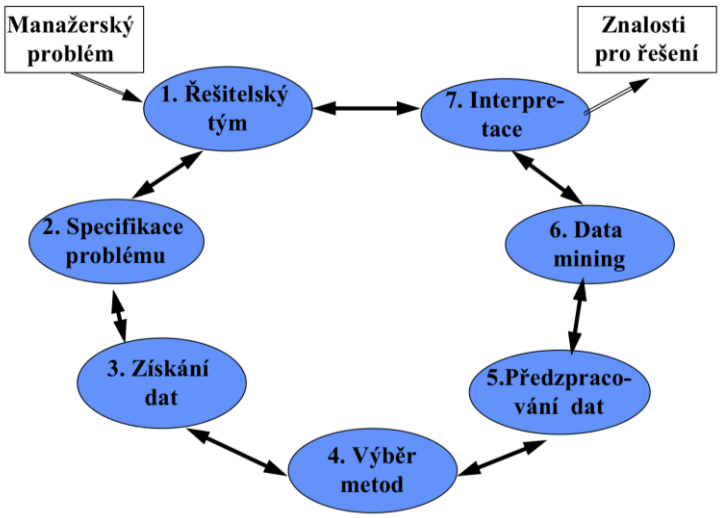
postprocessing. Na výstupu jsou pouze "užitečné" informace, ze kterých jsou následně získávány znalosti (Tan, 2018, s.3-4). Aby bylo možné získávat znalosti a poznat "užitečné" informace, nestačí jen definovat technologický proces, je třeba znát a určit bližší kontext věci a během celého *KDD* procesu s ním pracovat.



Obrázek 2: Proces KDD dle Fayyad et al. (Fayyad et al., 1996 podle Berka, 2003, s.16)

Brachman & Anand (Brachman & Anand, 1996 podle Fayyad, 1996, s.84) vnáší do *KDD* procesu kontext řešeného problému a vytváří devítikrokový proces s opakovanými iteracemi. P. Berka uvádí Anandův manažerský pohled (Anand, 1996 podle Berka, 2003, s.16-18) v sedmi krocích. Je vizuálně znázorněn jako kruh (viz Obrázek 3: Manažerský pohled na KDD), což již vizuálně evokuje jeho iterativní charakter. Určení problému na vstupu a znalost na výstupu jsou znázorněny mimo tento kruh, celkově tak proces obsahuje kroků devět. Na vstupu manažerského pohledu je byznysový problém, který je třeba vyřešit získáním co nejvíce relevantních informací, které mohou přispět k řešení. Obecným příkladem může být snaha o dosažení co největšího konverzního poměru. Marketingovou strategií pak může být nalezení skupiny uživatelů, kteří by na základě historie nákupů mohli mít o daný produkt větší zájem, než ostatní, a na tuto zacílit reklamu. Identifikace této skupiny uživatelů pak proběhne pomocí data miningových technik. V prvním kroku *KDD* procesu je třeba postavit tým sestávající z experta na řešenou problematiku, experta na data a experta na metody *KDD*. Jednotliví experti pracují společně, konzultují s externími experty a v případě potřeby mají k dispozici svůj vlastní tým. V následujícím kroku společně definují řešený problém z pohledu dobývání znalostí. Ve třetím kroku posuzují veškerá data, která jsou k dispozici, někdy uložena i v několika různých zdrojích včetně externích, a určí, která jsou podstatná pro řešení úlohy a jak je získat. Ve čtvrtém kroku je výběr vhodné metody,

resp. techniky. V případě identifikace určité skupiny uživatelů by byla jako technika vybrána např. metoda shlukování. V pátém kroku se vybraná data z druhého kroku upraví do vhodné podoby k aplikaci data miningové techniky. V šestém kroku dochází k aplikaci vybrané data minigové techniky, často i k opakované aplikaci v kombinaci s jinými vybranými technikami na základě dílčích výstupů. Výstupním krokem celého cyklu je interpretace, jejímž cílem je uvést do kontextu výsledky z jednotlivých iterací a určit, které výstupy jsou nezajímavé, které lze použít a které mohou být dány do podoby srozumitelné uživateli. Na základě interpretace a získaných znalostí jsou pak činěna opatření či se vypracuje zpráva.



Obrázek 3: Manažerský pohled na KDD (Anand a kol., 1996 podle Berka, 2003, s.16)

Fáze *Data mining* využívá algoritmické metody, které extrahují informace z dat. Cíle data miningu mohou být následující: ověření uživatelských hypotéz nebo objevování nových, dříve neznámých, skutečností. Dle Fayyada a kol. (Fayyad et al., 1996, s.85) většina data miningových technik vychází z několika ověřených metod strojového učení, rozpoznávání vzorů a statistiky. Těmi jsou klasifikace, shlukování, regrese a další. Podrobněji se data miningovým technikám věnuje kapitola 2.3. Dá se říci, že spousta data miningových technik je postavena na několika základních technikách, kdy každá z nich řeší některé problémy lépe, než ty ostatní. Univerzální pravidla však neexistují a volba nejvhodnější metody pro jednotlivé aplikace, aby bylo dosaženo stanovených cílů, bývá někdy výzvou. Množství dostupných algoritmů leckdy zmátne nejen začínajícího datového analytika, ale i toho zkušenějšího. Přitom vyjádření funkce každého takového modelu je

postaveno na kombinaci několika základních a velmi dobře známých funkcí, jako jsou polynomické funkce, booleovské funkce a další. V každém případě, dle Handa (Hand, 1994 podle Fayyad, 1996, s.86), Langlaye a Simona (Langlay & Simon, 1995 podle Fayyad, 1996, s.86), při neúspěšných opakovaných pokusech má větší význam upřesňování formulace problému než optimalizace detailu algoritmu. Z vlastní zkušenosti z praxe mohu toto tvrzení potvrdit, kdy v některých případech byl "chybný" výstup opraven až po doupřesnění podstatného detailu zadání, který předtím nebyl znám.

2.1.2. Uplatnění KDD

Uplatnění metod *Dobývání znalostí z databází* bylo využíváno v komerční sféře již od počátků oboru za účelem získávání konkurenčních výhod. Tato snaha, v kombinaci se značnou praktickou zkušeností získávanou v čase, pak na přelomu tisíciletí vedla k vytvoření několika metodologií, které jsou rozepsány v následující podkapitole. Veškeré podstatné úlohy, na kterých stojí prakticky celé odvětví *Dobývání znalostí z databází*, byly vyřešeny již před 30 lety. Již tenkrát byly využívány metody strojového učení, statistiky či neuronových sítí. V počátcích oboru byl hlavním limitem dostupnost výpočetního výkonu, avšak tento problém v čase postupně vymizel a v současnosti se nejklíčovější otázkou stala volba nejvhodnější data miningové techniky (Oreški, 2018, s.248). Dnes, díky několikanásobně většímu výpočetnímu výkonu a dostupnosti současných technologií, jsou i pokročilé metody aplikovány oproti minulosti celkem běžně, byť fundamentálně obtížnější úlohy a některé otázky staré 30 let stále vyřešeny nebyly. Oproti minulosti jsou navíc mnohem dostupnější znalosti a zkušenosti, které lze získat prakticky kdykoli online nejen přes MOOC platformy, ale i z např. bohatých diskusních fór.

Na první pohled se může zdát, že vývoj v čase přinesl jen samé pozitivní aspekty, avšak ruku v ruce s technologickými možnostmi a s prudkým nárůstem množství dat, se čím dál více prohlubuje termín *informačního přehlcení*. Informační přehlcení je stav zahlcení velkým množstvím informací, které mohou být nadbytečné, či nerelevantní a "způsobuje neschopnost vytěžit potřebné znalosti z nezměrného kvanta informací" (Sklenák a kol., 2001. s.5). Tento stav tak z manažerského pohledu přímo i nepřímo zasahuje i do oblasti *KDD*. Množství dat roste čtyřikrát rychleji, než světová ekonomika. Dle dostupných předpovědí (IDC, 2018) má v roce 2025, oproti současným 33 ZB v roce 2018, tvořit globální datasféru až 175 ZB dat. Bohužel i přes významný technologický posun lze v

současnosti z celého objemu dat zpracovat a vyhodnotit pouze 0,5%. Firmy sice vlastní a i nadále budou vlastnit stále se zvyšující velký objem dat, avšak s tímto množstvím zároveň vzniká i nejasnost v tom, co a jak vlastně vyhodnocovat. Vyhodnocovat "balast" je v současnosti v daném kontextu nejen velmi drahé, ale i matoucí. O to více je dnes důležité identifikovat, jaká data a informace jsou podstatná, aby znalost přinesla onu kýženou konkurenční výhodu, a jak získané informace předat ostatním a interpretovat je. Nehledě na prostředky, byť dobývání znalostí z databází v rozumném čase a za rozumnou cenu je jednou z priorit, lze technologicky realizovat prakticky cokoli. Ovšem výsledný drahý a přeplácaný dashboard plný čísel a grafů žádnou přidanou hodnotu nepřináší, ba naopak.

V důsledku informačního přehlcení a rychle měnícím se trendům ve vyhodnocování může dle mého názoru vznikat úskalí v tom, že se v praxi stane, že pro kýžené dokončení nejsou splněny již základní předpoklady v rámci manažerského pohledu v procesu dobývání znalostí z databází. Úskalí tkví v nesprávném nebo neúplném nadefinování řešeného problému a způsobu vyhodnocování. Následkem toho je výstup, který je formálně "správně", zároveň však neodpovídá na otázky problému.

Terminologie, kterou zavedl v 90. letech Fayyad a kolektiv, se v průběhu času rozšiřovala o další označení, která v současné době nemají ustálené jednoznačné významy, ovšem stále vychází z principu *KDD*. Pojmy *data science*, *data engineering*, *data stewardship, business intelligence* a další, nemají sjednocené a přesně vymezené definice a v různých interpretacích se významy pojmů překrývají či nahrazují. Paralelně vznikají i názvy pro nové pracovní pozice, jako je *data scientist*, *data engineer*, *data steward, data curator* a další, které rovněž nemají ustálené významy a náplně práce pod stejným označením se v různých firmách liší.

Globální technologické společnosti prezentují své vlastní rozdělení těchto pracovních rolí. Například Microsoft určil kombinaci pracovních rolí, které by měl mít každý úspěšný moderní datový projekt, pod označeními *data engineer*, *data scientist* a *AI (artificial intelligence) engineer* (Microsoft, 2019). Většinou má každá z pracovních rolí na starosti jinou fázi z hlediska *životního cyklu dat*, eventuelně pracují na úkolech paralelně. Z důvodu neukotvených definicí a současně při nedostatečné komunikaci se domnívám, že může docházet k duplikaci stejných činností v oddělení. V praxi se také může stát, že např. na pozici *data engineer* hledají člověka, který by měl zároveň umět vyhodnocovat kampaně

a prezentovat výsledky před vedením. Dle mého názoru, tyto negativní okolnosti můžou vznikat při nedostatečně uchopené koncepci během snahy o *digitální transformaci* firmy a *rozhodování na podkladě dat* ("data-driven decision"). Digitální transformace firem, nový pohled na propojení lidí, dat a procesů (Microsoft, 2019), je v současnosti velké téma, o kterém se "všude mluví" a nikdo nechce, aby jim "ujel vlak" a tím dle mého názoru vzniká hrozba uplatnění "rychlých řešení", ideálně za minimum peněz, v důsledku bez jasné koncepce, bez dlouhodobé strategie a s reálně nulovou přidanou hodnotou.

Další aktuální rizikovou skutečností může být, že s rostoucí popularitou fenoménu *KDD* se v komerční sféře stává, že v praxi dochází ke zkresleným představám o tom, co vlastně označuje. Domnívám se, že dnes, v kombinaci s populárními pojmy umělá inteligence a machine learning, může v praxi vznikat představa, že lze vyrobit nějaké "kouzelné tlačítko", které firmě vyřeší všechny její problémy a bude generovat trvalý zisk. Tyto mylné představy v kombinaci s nekonzistenčností dat a s představou, že analytik nebo dokonce stroj, najde odpovědi bez zapojení lidí z oboru, pak mají za následek to, že více než 85% takových projektů selže (Janča, 2018).

2.1.3. KDD v kontextu marketingu

V oblasti marketingu, za účelem lepší interakce se zákazníkem a pro tvorbu obchodních a kampaňových strategií, lze sledovat a vyhodnocovat spoustu dimenzí a metrik. Tyto nejsou ustálené a především v online marketingu se v čase dynamicky mění, kdy některé zanikají a místo nich vznikají nové. Tato dynamičnost kopíruje aktuální trendy ve vyhodnocování a trendy v oblasti zákaznické zkušenosti. Trend zákaznické zkušenosti, *CX*, se vyvíjí v důsledku rostoucího vlivu sociálních sítí na uživatele internetu a dále v důsledku stále větší dostupnosti informací a možnosti výběru z několika nabídek. Tomu je třeba přizpůsobit obchodní a kampaňové strategie. Nejčastěji užívanou data miningovou metodou v oblasti marketingu je klasifikace (Kaefera et al., 2005 podle Pavlovic et al, 2014).

V daném kontextu může být *KDD* nástrojem pro de-masifikaci kampaňových strategií či pro de-masifikaci komunikace se zákazníky. Pojem de-masifikace zavedl Alvin Toffler (Tai, 1997, s. 104). V minulosti běžné mass marketingové strategie jsou dnes na ústupu a nahradila je personalizace obsahu a osobní přístup k zákazníkovi. Mass marketingové strategie využívají masová média jako televize, rádio, časopisy a noviny k

předávání informace jak stávajícím, tak i potencionálním zákazníkům. Pro osobní přístup k zákazníkovi v rámci přímého marketingu jsou využívány kanály jako je například e-mail či telefon. Úspěch kampaní přímého marketingu závisí na dostupnosti zboží, způsobu komunikace, správnému načasování a výběru klientů (Pavlovic et al, 2014). Určení klíčových metrik marketingových kampaní pro jejich vyhodnocování je součástí strategie porozumět potencionálním i stávajícím zákazníkům a uspokojit jejich potřeby. Důvodem pro uplatnění této strategie je získat a udržet si loajálního zákazníka, který firmě přináší stabilní zisk, v důsledku dokonce větší, než zisk nově akvirovaného zákazníka. Hledat nové zákazníky stojí čas a peníze a z porovnání nákladů na masové versus cílené kampaně přímého marketingu vyplývá, že získat nového zákazníka je 12 krát dražší než si udržet stávajícího (Torkzadeh et al., 2006 podle Pavlovic et al, 2014). Navíc stávající zákazníci generují větší zisk a větší marži než noví zákazníci (Reichheld and Sasser, 1990 podle Pavlovic et al, 2014). Udržení si dlouhodobého zákazníka je dosaženo především pomocí personalizované interakce. Personalizované interakce lze dosáhnout například segmentací zákazníků na základě určitých kritérií a následně se zákazníkům v jednotlivých segmentech co nejvíce přiblížit. Pro personalizaci a tvorbu marketingových strategií je nejčastěji využíváno nástrojů *CRM* (Customer Relationship Management) (Reinartz and Kumar, 2002; Torkzadeh et al., 2006 podle Pavlovic et al, 2014), pomocí kterých jsou shromažďována a ukládána data o chování a charakteristikách jednotlivých zákazníků pro následné marketingové strategie. Data o chování zákazníků jsou pro budování dlouhodobých vztahů klíčová (Coussement et al., 2009 podle Pavlovic et al, 2014). Dnes si společnosti uvědomují, že stávající zákazník uložený v jejich databázi je jejich nejcennějším aktivem (Athanassopoulos, 2010 podle Pavlovic et al, 2014). Data miningové techniky umožňují nalézt souvislosti a vzory v rozsáhlých databázích se zákaznickými daty a tyto výstupy pak použít v rámci personalizace. Vedle udržení si stávajících zákazníků lze s využitím prediktivních technik identifikovat nové uživatele, kteří s velkou pravděpodobností provedou konverzi.

V současnosti využívané dimenze a metriky pro vyhodnocování dle aktuálních trendů můžou již za několik měsíců být zastaralé. Určení aktuálních klíčových marketingových identifikátorů (*KPI*), na základě kterých lze optimalizovat a stavět další kampaňové strategie, tak vyžaduje neustálé vzdělávání v oboru. V souvislosti s *KDD* procesem je zároveň důležité, aby každá taková změna a optimalizace byla v rozumném čase technicky proveditelná, proto efektivní spolupráce lidí, zastřešující technické řešení, a lidí,

kteří určují marketingová *KPI* a na základě znalostí získaných z dat vydávají strategická rozhodnutí, má čím dál větší váhu.

Při vyhodnocování marketingových kampaní je třeba rozlišovat kampaně brandové a výkonnostní. Data mohou být kampaňová či webová, která měří chování uživatelů na webu. Kampaňová data jsou data online kampaní, které jsou spuštěny na různých online kanálech. V závislosti na typu kampaně (brand, výkon) se vyhodnocují metriky typické pro daný typ kampaně. U brandových kampaní je klíčové zvýšit povědomí o značce a mít co největší zásah reklamy. Cílem je, aby se o tom, co kampaň propaguje, dozvědělo co nejvíce lidí. Naopak výkonnostní kampaně jsou zaměřené na splnění konkrétního cíle a vyhodnocuje se především počet splnění cíle (například počet vyplnění formuláře, počet objednávek a další). Brandové a výkonnostní kampaně se vyhodnocují pomocí různých metrik. Měří se počet zobrazení banneru či videa, počet kliknutí na jednotlivé bannery, zhlédnutí videa, počet konverzí. Počet konverzí je počet provedení akcí, které měly být cílem reklamní kampaně (například vložení zboží do košíku, vyplnění formuláře atd.). Vedle výše zmíněných metrik se vyhodnocují počítané metriky, nejčastěji míra prokliku (CTR(%) = počet kliků/počet zobrazení reklamy), cena za kliknutí (CPC = cena/počet kliknutí), cena za akci - konverzi (CPA = cena/počet konverzí), či konverzní poměr (CR(%) = počet konverzí/počet návštěv). Data lze pro vyhodnocení agregovat na základě umístění reklamy, média, kampaně, kanálu či časového úseku.

Data webové analytiky měří návštěvy a chování uživatelů na webu, sledují zdroj, odkud uživatel přišel - zda z placeného zdroje (kampaň) či neplaceného zdroje (organické vyhledávání či organická návštěva webu). Nejčastěji je k měření dat webové analytiky využívána platforma Google Analytics. Pomocí webové analytiky lze měřit a vyhodnocovat počet návštěv jednotlivých stránek webu, počet unikátních návštěv, počet konverzí, konverzní poměr, čas strávený na stránce, počet navštívených stránek za jednu návštěvu, tržby a další metriky. Dimenzí může být například vstupní či jakákoli jiná stránka, zdroj návštěvy atd. Pomocí Google Analytics lze měřit nejen data z webů, ale dávat je do kontextu s kampaňovými daty a s transakčními daty z e-shopů. Pomocí transakčních dat a dat monitorující chování zákazníka na webu e-shopu lze sledovat celou cestu zákazníka od první návštěvy webu až po případné dokončení nákupu včetně posloupnosti kroků.

Z výše uvedené podkapitoly vyplývá, že data mining nelze nikdy vydělit jako nezávislou disciplínu. Vždy navazuje na fázi před, předzpracování dat, a na fázi po, interpretaci. Tyto fáze tvoří celek procesu, kdy při vynechání jedné z fází by nebylo možné získat validní a relevantní znalost. Například při vynechání očištění dat v rámci jejich přípravy může být výstup chybný, jelikož bude zahrnovat nerelevantní vstupy. Nebo při nevhodné volbě formy výstupu, například zavádějící vizualizaci, mohou vznikat obtíže pro správnou interpretaci. Vývoj oboru a paralelní růst výpočetního výkonu v kombinaci se snadnou dostupností k informacím přináší nejen pozitivní dopad, ale možná i související úskalí. V oblasti marketingu je v současné době *KDD* proces v důsledku de-masifikace kampaňových strategií a de-masifikace komunikace se zákazníky především nástrojem pro personalizaci reklamy a komunikaci se zákazníky.

2.2. Standardy pro dobývání znalostí

Hledání optimálního procesu pro *Dobývání znalostí z databází* bylo přítomno již od počátků oboru v průběhu 90. let minulého století. V této souvislosti postupně vznikaly různé komerční i univerzální metodologie, jejichž cílem bylo předat ostatním nabyté zkušenosti. Metodologie pro dobývání znalostí z databází zrcadlí hierarchické schéma data - informace - znalosti a většinou vychází z původní Fayyadovy definice *KDD* procesu z roku 1996 (viz Obrázek 1: Vývoj metodologií). Známou metodologií od producentů, která vychází z *KDD* procesu, je metodologie *SEMMA* od firmy SAS. Další známou metodologií od producentů je metodologie *5A* od firmy SPSS. Firma SPSS byla také součástí konsorcia firem, kteří se podíleli na vývoji pozdější metodologie *CRISP-DM*. *CRISP-DM* vychází jak z těchto dvou zmiňovaných, tak i z dalších několika metodologií, které vychází z *KDD* procesu. Technologicky a oborově univerzální metodologie *CRISP-DM* odráží několikaleté zkušenosti autorů z různých prostředí byznysu, aplikací a používaných technologií a je jakýmsi průvodcem při data miningových projektech. Úspěch i dnes stále nejrozšířenější metodologie tkví právě v praxi a ve zkušenostech autorů, kteří během jejího vzniku pracovali na širokém spektru projektů. Přestože většina metodologií pro dobývání znalostí z databází vychází z původního Fayyadova *KDD* procesu, jehož základní logiku přebírají všechny, lze mezi nimi nalézt rozdíly ve způsobu rozdělení a označení jednotlivých fází či rozdíly v tom, zda je na začátku přítomna byznysová analýza řešeného problému či ne, či rozdíly v tom, na jakém základě proběhne výběr dat k dalším krokům.

2.2.1. Metodologie pro dobývání znalostí od producentů

Nejznámějšími metodologiemi od producentů jsou metodologie *5A* a metodologie *SEMMA*. Metodologie *SEMMA* má přímou relevanci k Fayyadovu *KDD* procesu, zatímco metodologie *5A* z ní sice přímo nevychází (viz Obrázek 1: Vývoj metodologií), avšak logická posloupnost kroků je totožná. Společným jmenovatelem metodologií je vždy dodržení hierarchického schématu data - informace - znalosti.

Metodologie *SEMMA* byla Paulem Chikem z firmy SAS Institute představena na osmém mezinárodním workshopu *Data Mining, Data Warehousing & Client/Server Databases* v Hong Kongu v roce 1997, jako doporučený postup při data miningových projektech s využitím jejich softwarového produktu Enterprise Miner. Označení *Data mining* zde významově odpovídá Fayyadově označení *KDD*. Proces tvoří několik kroků, kdy na začátku je vždy definice řešeného problému a až následně probíhá volba způsobu, jak na konkrétní problém nalézt odpověď. *Data mining* dle jejich definice označuje pokročilé metody pro zkoumání a modelování vztahů mezi velkým množstvím dat. Jedná se o technologii, ne o předpřipravené řešení byznysových otázek. Data mining jako takový nemůže být nikdy plně automatizovaný, jelikož výstup z něj je vždy plně závislý na konkrétním vstupu. K řešení tedy nelze přistoupit aplikací jedné či dvou technik na veškeré byznysové otázky (Chik, 1997). Každé řešení konkrétního problému je tak svým vstupem individuální. Metodologie *SEMMA* je určitým "přehledem" *KDD* procesu, má iterativní charakter a skládá se z pěti kroků, jejichž označení je akronymem pro název metodologie. Obsahuje kroky Sample (vzorek), Explore (prozkoumat), Modify (modifikovat), Model (modelovat), Assess (posoudit).

Krok Sample zahrnuje výběr vzorku dat, který má být dostatečně velký, aby obsahoval veškeré relevantní informace, ale zároveň malý pro rychlou manipulaci s daty. Až v dalším kroku Explore dochází k prozkoumání daného vzorku dat, zda obsahuje určité anomálie či na první pohled zjevné trendy. Lze k tomu využít jak vizualizace dat, tak i statistické analýzy, ale pouze na vybraném vzorku dat. Ačkoli z hlediska výpočetního výkonu té doby je tento postup pochopitelný, osobně zde vnímám určitý rozpor, kdy bez předchozí znalosti dat, dle mého názoru, nelze s jistotou vybrat vzorek dat, který je dostatečně velký z hlediska obsahu, ale zároveň malý z hlediska výkonu. Samozřejmě záleží na úhlu pohledu a typu úlohy, jak krok Sample pojmout. V příkladu odhalování podvodného užití kreditních karet, který Chik na osmém mezinárodním workshopu během prezentace

metodologie prezentoval, je krok výběru vzorku dat pojat typicky jako výběr dat pro úlohy strojového učení. Ze znalosti, že celkový poměr ukradených a neukradených karet je v poměru přibližně 1:5, vybral náhodně vzorek dat 20 000 ukradených karet a 100 000 neukradených karet. Zde tato posloupnost kroků nevadí. Avšak v případě jiných úloh, u kterých jsou podstatné určité trendy či nepředvídatelné anomálie v datech (například v rámci sledování množství výskytů zákazníků s určitými charakteristikamii), takto nahrubo data vybrat nelze a výběru vzorku dat by mělo předcházet jejich důkladné prozkoumání, byť někdy i za cenu většího výpočetního výkonu. Je nutné zmínit, že v roce 1997, kdy tato metodologie byla na workshopu představena, byly možnosti výpočetního výkonu řádově jiné než dnes. S ohledem na tuto skutečnost by v dnešní době prozkoumání dat před výběrem vzorku nemělo vyžadovat takové výpočetní nároky, jako v minulosti, avšak s rostoucím množstvím dat by finanční náklady na tento postup mohly být i dnes neúměrné (např. prozkoumání Big Data databáze pro IoT).

V následujícím kroku Modify, poté, co víme, jaká data jsou k dispozici, přichází na řadu jejich modifikace pro konkrétní modelovací techniky v dalším kroku. Modifikací je míněno např. výběr relevantních dat, čištění dat, vytvoření nových veličin pro požadované výpočty, transformace dat a další. Iterativní charakter metodologie umožňuje data opakovaně modifikovat při zjištění nových skutečností. V kroku Model jsou aplikovány konkrétní data miningové techniky, jako jsou např. neuronové sítě, modely založené na stromech, logistické modely či další statistické modely, na data připravená v předchozím kroku. Výběr data miningové techniky závisí na řešeném problému a na limitech dat, která jsou k dispozici. Krok 5, Assess, posuzuje užitečnost a validnost výstupů z kroku Model. Úspěšnost a užitečnost výstupu musí být ověřena jak na vzorku dat pro validaci, tak i na v modelu použitých tréninkových datech. Test modelu na datech, u kterých známe výsledek, zároveň určí jeho přesnost a validnost. Po zhodnocení modelu, v případě získání odpovědí, která jsou relevantní a přináší řešitelům požadovanou znalost, zároveň otestované na dostatečnou přesnost a validnost, lze mluvit o úspěšném a užitečném řešení, které vede k novým znalostem a tvorbě strategie. Z tohoto důvodu by měl být finální výstup srozumitelný širokému spektru uživatelů. V případě neúspěšného výstupu, iterativní charakter metodologie umožňuje vrátit se k předchozím krokům a jednotlivé výstupy ladit a zpřesňovat, včetně úprav prvotního výběru dat. Nástroj Enterprise Miner je k dispozici i v současnosti a dle aktuální dokumentace (SAS, 2017) využívá původní, nijak nezměněnou, metodologii *SEMMA*. Dle mého osobního názoru by prospěla revize kroků Sample a Explore

vzhledem k aktuálním výpočetním možnostem. Rozhraní platformy je dnes plně přizpůsobeno nejčastějšímu koncovému uživateli - byznys analytikovi s minimem statistických znalostí, kterého nástroj celým procesem intuitivně provede a umožní mu vyzkoušet a porovnat úspěšnost několika technik. Pokročilejší uživatel může využít i rozšířených možností. Tím, že je metodologie navázána na aktuální komerční produkt, je zřejmě důvod, že využívání metodologie je hned na třetím místě za metodologií *CRISP-DM* a za vlastními metodologiemi uživatelů, byť její využití od roku 2007 kleslo o několik procentních bodů (Piatetsky-Shapiro, 2014).

Název metodologie *5A* od firmy SPSS je taktéž akronymem pro jednotlivé kroky procesu, stejně jako název *SEMMA*. Oproti metodologii *SEMMA* je tato v dnešní době využívána spíše minimálně - v anketě (Piatetsky-Shapiro, 2014) je zřejmě pouze součástí možnosti "Ostatní, doménově nespecifikované". Metodologie *5A* je dle mého názoru více prakticky zaměřená už jenom tím, že je kladen důraz nejen na byznysové zadání, ale i na důkladné pochopení problematiky řešeného oboru, a je doporučována automatizace úspěšných řešení. Doporučení automatizace vnímám na základě svých dosavadních zkušeností jako důležitý krok při zavádění data-driven procesů do firmy. Název *5A* označuje, stejně jako *SEMMA*, pět kroků procesu, a to Asses (posouzení a stanovení cílů), Acces (získání potřebných dat), Analyze (analýza dat), Akt (přeměna výstupů analýz na akční znalosti), Automate (automatizace procesu u opakujících se úloh) (SPSS Metoda 5A podle Berka, 2003, s.23). Stejně jako u předchozích metodologií, je zde na pozadí hierarchické schéma data - informace - znalosti, avšak pod jinými označeními. Shodně v jednotlivých krocích pro získání informací jsou postupně dávány do kontextu data a následně dávány do kontextu získané informace, pro získání znalostí. U metodologie *5A*, pro schopnost relevantního určení problému a otázky jako takové, je v prvním kroku kladen velký důraz na pochopení problematiky oboru, ve kterém se daný problém řeší, a znalost odborných dat, ze kterých lze odpovědi vyčíst. Například v dynamickém prostředí dimenzí a metrik online marketingu může mylné pochopení problematiky a dat přinášet řádově chybné výstupy už chybným určením vstupu. Problematické by pak bylo, pokud se z neznalosti oboru výstup z nesprávně zvoleného vstupu označí za správný a v rámci posledního kroku se model zautomatizuje. Ve druhém kroku metodologie je získání potřebných dat, kterému předchází důkladná analýza veškerých dostupných relevantních zdrojů. Obsahem třetího kroku Analyze je užití různých Data miningových technik, porovnání jejich výsledků, úspěšnosti a volba té nejjednodušší, nejpřesnější a zároveň validní techniky. Během čtvrtého kroku Akt

se převádí výstupy z předchozího kroku, které musí být v jasné a srozumitelné podobě, do akčních znalostí, např. do manažerských rozhodnutí. V posledním kroku Automate je doporučení úspěšné modely zautomatizovat, a to ty, jejichž výstupy jsou často využívány a to nejen pro jejich sledování, ale i pro sledování důsledků rozhodnutí v souvislosti s získanými akčními znalostmi (SPSS Metoda 5A podle Berka, 2003, s.23)

Obě metodologie mají shodný počet kroků a podobné logické schéma na pozadí, přesto se liší v detailech a to nejen z důvodu návaznosti na různé softwarového nástroje, ale i z důvodu různé firemní konvence přístupu k řešení. Na Obrázku 1 (Vývoj metodologií) jsou vidět i další metodologie, které následovaly Fayyadův *KDD* proces, ať už z něj vycházely, či ne. V detailech nejednotnost všech dílčích metodologií o pár let později vyřešila dnes stále nejrozšířenější a dodnes nepřekonaná metodologie *CRISP-DM*, která se opírá nejen o zkušenosti a praxi autorů, ale inspiruje se i u předešlých metodologií, z nichž dvě od producentů, *SEMMA* a *5A*, byly v kapitole rozebrány.

2.2.2. Metodologie CRISP-DM

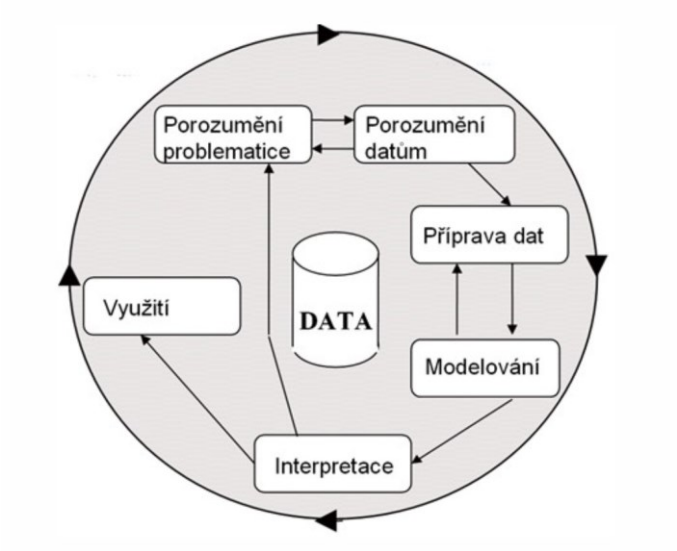
Ve druhé polovině devadesátých let vzniklo konsorcium pod akronymem *Cross-Industry Standard Process for Data Mining (CRISP-DM)*. Metodologie nabízela souhrn znalostí a zkušeností širokého spektra uživatelů nezávisle na oblasti byznysu, technologiích či aplikaci. Tato metodologie, *CRISP-DM*, je i v současnosti, dle Gregory Piatetsky (Piatetsky-Shapiro, 2014), stále nejpoužívanější a opírá se o ni jak vědecká oblast, tak i komerční. Dokládá to také množství vědeckých prací na téma dobývání znalostí z databází, ze kterých jsem v diplomové práci čerpala, a které se o metodologii *CRISP-DM* opírají. Metodologie *CRISP-DM* strukturou a logikou vychází z kombinace technologického pohledu na proces *KDD*, který nastínil Fayyad a kol. v roce 1996 a z manažerského pohledu Ananda a kol., který vznikl v témže v roce. Ačkoli jsem se ve své dosavadní praxi ani jednu z výše zmiňovaných metodologií při práci nevyužívala, nakonec jsem po získání určitých znalostí a zkušeností, přirozeně a intuitivním přístupem, došla k vlastní, která je ze všech zmíněných metodologií nejblíže právě metodologii *CRISP-DM*, byť samozřejmě nikdy nepokryla její komplexnost. I to je důvodem, proč jsem si tuto metodologii vybrala v rámci aplikace praktické části diplomové práce.

To, že již skoro 20 let, je metodologie *CRISP-DM* stále nejvyužívanější a nejúspěšnější ze všech, připisuji především tomu, a i jak sami autoři v úvodu dokumentace k metodologii *CRISP-DM* uvádí, metodologie nebyla vytvořena na čistě teoretické akademické půdě, či za zavřenými dveřmi úzkou skupinou odborníků, ale vznikla na bázi sdílení praktických znalostí a zkušeností široké skupiny lidí, kteří pracovali na spoustu reálných projektů v různých oborech s využitím různých technologií. Přestože se projekty, označením fází a terminologií, v detailech lišily, nadřazený pohled na proces data miningu byl u všech vždy stejný. To bylo pro autory metodologie *CRISP-DM* během workshopu, který v souvislosti s uchopením metodologie uspořádali, klíčovým zjištěním, a zároveň konfirmací, že univerzální metodologii vytvořit lze (Chapman et al., 2000). Metodologie *CRISP-DM* reflektuje poznání založené na sdílené praxi z reálného světa, detailně popisuje a větví základní logiku, kterou lze aplikovat na jakýkoli projekt a technologii, a tímto se stává nezávislou a univerzálně použitelnou.

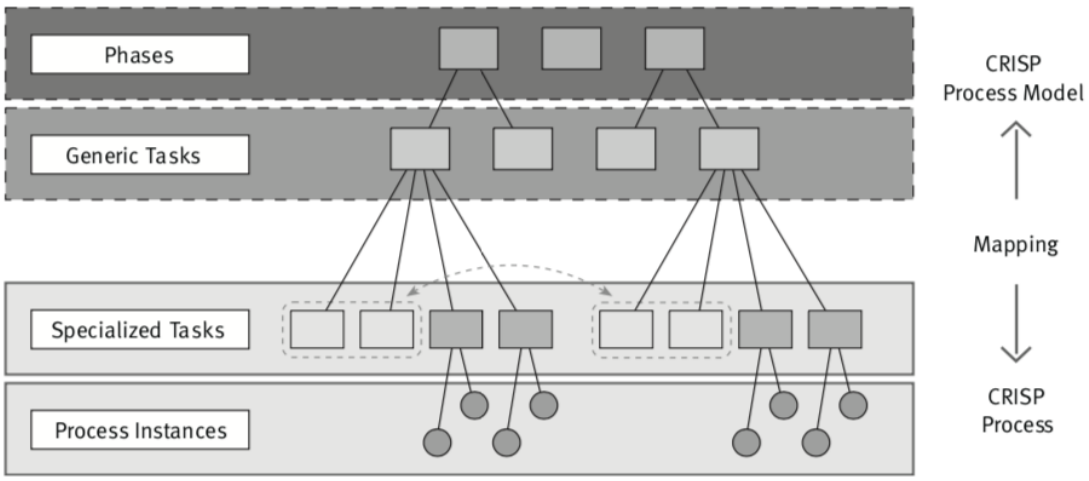
Metodologie má čtyřúrovňovou hierarchickou strukturu, která pokrývá jak obecný model procesu na první úrovni, tak i jeho konkrétní procesy na úrovni čtvrté (Viz Obrázek 5: Hierarchie úrovní metodologie *CRISP-DM* v angličtině). První a nejobecnější úroveň tvoří šest fází, které rámcově odpovídají krokům *KDD* procesu či krokům metodologie *SEMMA* či *5A* a mají stejně jako tyto iterativní charakter s neomezeným množstvím opakování (Viz Obrázek 4: Fáze metodologie *CRISP-DM*). V první fázi metodologie *CRISP-DM*, na začátku celého procesu, je klíčové porozumění řešené problematice, cílům a požadavkům z obchodního hlediska a jejich následné převedení do konkrétního zadání a cílů z hlediska *KDD* a vytvoření předběžného návrhu plnění. Druhá fáze, porozumění datům, zahrnuje definici dat potřebných ke splnění zadání a jejich získání. V této fázi se hodnotí kvalita získaných dat, analyzují se první poznatky o datech a mohou být již zjevné první informace, které lze z dat vyčíst. Na základě výskytů zajímavých podmnožin v datech lze vytvářet hypotézy o informacích skrytých.

Třetí fáze, příprava dat, zahrnuje veškeré kroky potřebné pro vytvoření datasetu, který bude použit při aplikaci data miningové techniky v další fázi modelování. Během přípravy dat ve třetí fázi se data čistí, transformují, redukuje jejich množství či se vytváří pomocné metriky. Předpokládá se, že mezi těmito dvěma fázemi, přípravou dat a modelováním, bude docházet k velkému množství iterací, jelikož jeden typ problému lze

řešit více data miningovými technikami a tyto jednotlivé techniky mohou vyžadovat různé datasety, ať už z hlediska formy, tak i obsahu.



Obrázek 4: Fáze metodologie CRISP-DM (Chapman et al., 2000, s.10)



Obrázek 5: Hierarchie úrovní metodologie CRISP-DM v angličtině (Chapman et al., 2000, s.6)

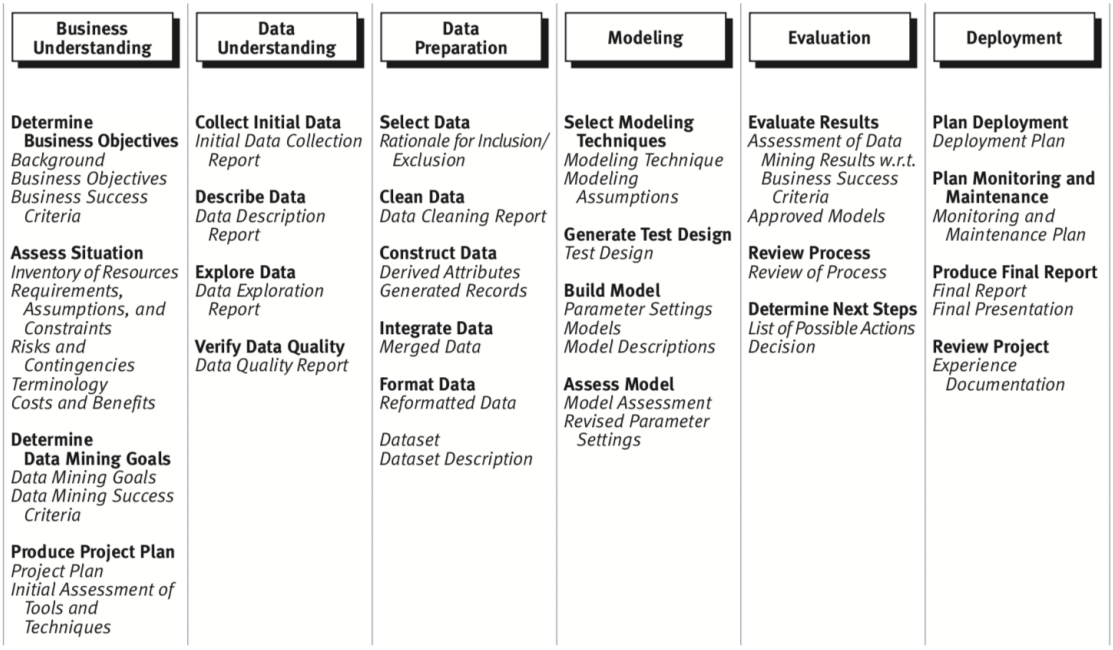
Během vyhodnocení v páté fázi, interpretaci, je třeba zhodnotit, zda výstup, který je sám o sobě z datového pohledu již hodnotný, má hodnotu i z obchodního hlediska a relevantně a úplně pokrývá problematiku definovanou v zadání. Během tohoto kroku je učiněno rozhodnutí, zda bude tento výstup využit v praxi, tedy zda postoupí do šesté fáze a přistoupí se k jeho využití, nebo zda bude vstupem pro novou iteraci. V šesté fázi se výstupy předávají zadavateli v pro něj srozumitelné podobě a klade se důraz především na porozumění zadavatele, co obnáší implementace modelu do praxe, jak může s výstupem pracovat a jaký užitek z něj může mít. Demonstrace zadavateli probíhá na praktických ukázkách reálného využití.

Ve druhé úrovni metodologie jsou popsány v obecné rovině veškeré kroky pro každou z fází, které je třeba vykonat u jakéhokoli data miningového projektu. Přidaná hodnota metodologie *CRISP-DM* je právě tato úroveň a další dvě podúrovně, které základní fáze procesu rozvětvují až do úrovně jednotlivých procesních kroků, které zahrnují veškeré dílčí vstupy a výstupy vznikající během celého procesu (viz Obrázek 6: Obecné úkoly metodologie CRISP-DM v angličtině). Druhá úroveň metodologie rozvětvuje jednotlivé fáze do dílčích kroků (na Obrázku 6: Obecné úkoly metodologie CRISP-DM označeno tučně), které obsahují seznam výstupů (na Obrázku 6: Obecné úkoly metodologie CRISP-DM označeno kurzívou), což je úroveň třetí. Pro jednotlivé výstupy jsou v rámci metodologie určeny konkrétní úkoly, které je třeba pro dosažení výstupů vykonat, což je úroveň čtvrtá. Příkladem je ve fázi modelování, což je označení první úrovně, vytvoření modelu, což je detail druhé úrovně, určení parametrů, což je úroveň třetí, a konkrétní procesní úkony jako je určení vstupních parametrů a dokumetace důvodů pro jejich volbu, což odpovídá úrovni čtvrté. Na čtvrté úrovni jsou popsány nejen tyto jednotlivé procesní úkony, ale i relevantní doporučení z praxe, ať už formou tipů či varování.

Obrázek 6 (Obecné úkoly metodologie CRISP-DM v angličtině) znázorňuje fáze a kroky, které lze popsat v češtině následovně. Ve fázi porozumění problematice jsou stanoveny cíle z obchodního hlediska, které zahrnují popis byznysu a jeho cílů a určení kritérií pro zhodnocení úspěšnosti z obchodního hlediska. V kroku posouzení situace jsou posuzovány dostupné zdroje, požadavky, předpoklady, omezení, rizika, včetně posouzení nákladů a přínosů. V následujícím kroku jsou stanoveny cíle z hlediska data miningu, jsou stanovena kritéria hodnocení. V závěru fáze je vypracován projektový plán.

Ve fázi porozumění datům jsou nejprve shromážděna iniciační data, ze kterých lze zjistit prvotní informace o charakteru dat. V následujících krocích jsou data prozkoumána a popsána jak z formálního hlediska, tak i z hlediska obsahu. Na závěr fáze je na základě předchozích kroků ověřena kvalita dat. Ve fázi přípravy dat probíhá příprava pro fázi modelování. Probíhá vhodný výběr dat pro konkrétní modelovací techniku a jejich čištění. Je vytvořen relevantní dataset, který splňuje jak formu, tak i obsah, pro aplikaci modelovací techniky v další fázi.

Ve fázi modelování je vybrána modelovací technika a vytvořen její návrh, na základě kterého je model postaven. V závěru fáze je dle získaných výstupů model posouzen. Součástí fáze interpretace je zhodnocení výsledků, posouzení procesu a určení následujích kroků. V případě rozhodnutí uplatnění výstupů v praxi se ve fázi využití vypracuje plán nasazení, monitorování a správy. Součástí fáze využití je i závěrečná fáze shrnující celý proces pro zadavatele a shrnutí projektu, obsahující zkušenosti a doporučení získané v dané iteraci.



Obrázek 6: Obecné úkoly metodologie CRISP-DM v angličtině. Tučně jsou označeny vstupy a kurzívou výstupy (Chapman et al., 2000, s.12)

Na první pohled je zjevné, že základní logika procesu metodologie *CRISP-DM* odpovídá jak Fayyadovu *KDD* procesu, tak i metodologiím od producentů. Všechny výše zmíněné metodologie zrcadlí hierarchické schéma data - informace - znalosti. Metodologie

od producentů a metodologie *CRISP-DM* se navzájem liší pouze ve způsobu rozdělení a označení jednotlivých fází. Dalšími rozdíly je přítomnost či absence počáteční byznysové analýzy a výběr dat k dalším krokům.

2.3. Vybrané KDD techniky vhodné v marketingové oblasti

Součástí metodologie CRISP-DM je výčet nejčastějších typů úloh v rámci dobývání znalostí z databází a doporučené způsoby jejich řešení pomocí různých data miningových technik. Mezi nejčastější typy úloh dle autorů metodologie, a které jsou dle mého názoru zároveň vhodné pro užití v oblasti marketingu, patří úlohy deskripce a agregace dat, úlohy segmentace, úlohy klasifikace a úlohy prediktivních analýz. V této kapitole jsou rozepsány vhodné techniky pro řešení těchto vybraných úloh, které jsou zároveň dány do kontextu s oblastí marketingu. Kromě úloh deskripce a agregace dat, které jsou součástí ranných fází metodologie *CRISP-DM*, jsou popsané úlohy a techniky součástí čtvrté fáze modelování. Z pohledu *KDD* procesu se jedná o krok *Dolování*, z manažerského KDD pohledu se jedná o krok *Data mining*.

2.3.1. Techniky deskripce a agregace dat

Deskripce a agregace dat je úloha, která je součástí ranných fází procesu Dobývání znalostí z databází a vede k popisu a porozumění datům. Uživatel za pomocí popisné statistiky a vizualizačních technik získá přehled o datech a definuje hypotézy o informacích skrytých, například při nalezení zajímavých segmentů v datech. Pro tyto úlohy je typické užití agregačních funkcí jako je součet, počet či aritmetický průměr. V některých případech však výstupy z úloh tohoto typu mohou být zároveň i výstupy celého procesu *KDD* (Chapman et al., 2000,). Výstupem celého procesu může být například porovnání počtu impresí za různá časová období dle typu kampaně nebo vizualizace počtu zákazníků dle věku a zájmů. Zadavateli je pak prezentováno finální shrnutí a jeho popis v pro něj srozumitelné podobě, například pomocí prezentace. V případě, že tento typ úlohy je zároveň i cílem celého *KDD* procesu, tedy pokrývá zadání z obchodního hlediska, lze k řešení využít běžné reportingové nástroje, u nichž nehraje roli, že nenabízí možnosti pokročilejšího modelování dat. Techniky deskripce a vizualizace dat lze také využít k detekci a analýzám anomálií v datech, které by mohly zkreslit výstupy celého *KDD* procesu, pokud by nebyly detekovány.

Popisná statistika zkoumá zákonitosti projevující se u velkého počtu prvků, jejichž souhrn je označován jako *data*. Data převádí do formy grafů a tabulek a vypočítává jejich číselné charakteristiky. Proměnné, neboli veličiny, jsou měnící se vlastnosti jednotek dat a jsou buď kvantitativní - mající numerickou hodnotu nebo kvalitativní - mající nenumerickou hodnotu (Průcha, 2010). V oblasti zpracování dat se lze setkat s označeními dimenze a metriky, kdy dimenze odpovídá kvalitativní veličině a metrika kvantitativní veličině. Dimenzí může být časová veličina, název kampaně či jméno zákazníka. Metrikou jsou například imprese (zobrazení), kliky či konverze. Počítané metriky jsou metriky získané výpočtem. Kvantitativní veličiny mohou nabývat diskrétních či spojitých hodnot. Diskrétní hodnoty jsou hodnoty ze zadané konečné množiny a spojité hodnoty jsou hodnoty ze zadaného intervalu (Průcha, 2010).

Aby vynikly charakteristické vlastnosti jednotek dat, je třeba je uspořádat tříděním. Třídění dle jedné veličiny je jednostupňové a třídění dle více veličin vícestupňové. Třídou (class) u kvantitativních dat je část dat, která patří do intervalu mezi největší (horní hranici) a nejnižší (dolní hranici) hodnotu třídy. Střed třídy (class mark) je průměr horní a dolní hranice třídy, šířka třídy (class width) je rozdíl mezi horní a dolní hranicí třídy. Absolutní četnost (frequency) je počet prvků (jednotlivých výskytů), které patří do určité třídy a relativní četnost (relative frequency) je poměr četnosti prvků třídy ku celkovému počtu prvků. Při třídění kvalitativních hodnot patří do stejné třídy prvky mající stejný znak nebo skupinu znaků (Průcha, 2010). Do stejné třídy, například do třídy brandových kampaní, se zařadí pouze ty kampaně, které mají v názvu řetězec znaků “brand”.

Pro znázornění charakteristiky dat lze využít tabulky nebo jejich vizualizaci. Ve své dosavadní praxi v marketingové oblasti jsem nejčastěji využívala tabulek a těchto několik typů grafů: sloupcový diagram, histogram, bodový graf, polygon četností a relativních četností, výsečový graf. Tabulka je tvořena kvalitativními veličinami a to jednou či více dimenzemi v různých úrovních granularity a k nim jsou přiřazeny kvantitativní veličiny, metriky. Dle míry detailu, kterou dimenze reprezentuje, lze kvantitativní veličiny zobrazovat od jednotlivých výskytů až po agregované hodnoty. Sloupcový graf tvoří oddělené sloupce a používá se především u vizualizací kvalitativních hodnot. Sloupce na ose x (horizontální ose) rezprezentují třídy a spojité hodnoty na ose y (vertikální ose) vyjadřují četnosti prvků těchto tříd. Na ose y (vertikální ose) lze ve sloupcovém grafu vykreslit i jinak agregované

hodnoty jako je například suma či průměrná hodnota veličin v rámci třídy. V případě přidání další kvalitativní dimenze do vizualizace lze provést rozlišení barvami sloupců.

Histogram je používán pro znázornění distribuce dat, kdy na ose x jsou měřenné veličiny a na ose y četnosti měřenných veličin. Má na první pohled podobné vizuální vyjádření jako sloupcový graf, ale oproti sloupcovému grafu jsou nejen na ose y, ale i na ose x, spojité hodnoty. Z tohoto důvodu se sloupce navzájem dotýkají. Na ose x (horizontální ose) jsou znázorněny jednotlivé výskyty prvků. Na ose y (vertikální ose) jsou znázorněné četnosti prvků. Polygon četností a relativních četností, neboli spojnicový graf, je graf obsahující jednu nebo více spojnic, což jsou body spojené úsečkami. Graf má obdobný princip jako sloupcový graf, ale je vykreslen pomocí spojnic. Na ose x je většinou umístěná časová řada. Více spojnic vznikne při vizualizaci více dimenzí, které je vhodné pro přehlednost odlišit různými barvami nebo typem čáry.

Bodový graf lze využít několika způsoby. Lze jej využít pro vykreslení podobným způsobem jako sloupcový graf, ale vizualizací body. V případě přidání další kvalitativní dimenze do vizualizace lze provést rozlišení různým tvarem či různými velikostmi bodů. Variantou bodového grafu je XY bodový graf vhodný pro korelační analýzy. Cílem korelační analýzy je zjistit, zda mezi dvěma numerickýma veličinami platí linerání závislost (Berka, 2003, s.49). Na obou osách jsou vynesené spojité kvantitativní hodnoty. Průsečíky těchto kvantitativních hodnot jsou označeny body. Dle hustoty bodů se určí vzájemný vztah mezi veličinami (korelace) - čím větší hustota bodů, tím silnější je vztah. Oblastí s největší hustotou bodů vede přímka vyjadřující směr korelace. Dvě veličiny jsou buď závislé nebo nezávislé, ale většinou má jedna z veličin vliv na druhou. V marketingu by se tento graf dal využít například v případě, že bychom měli k dispozici údaje o odhadovaných měsíčních příjmech zákazníků a hledali bychom korelaci s hodnotami cen produktů na e-shopu. Výsečový graf, označován též jako kruhový graf, znázorňuje velikostí výseče četností tříd. Každé třídě odpovídá jedna výseč (Průcha, 2010). Například u celkového počtu konverzí lze výsečovým grafem znázorit, jaké kampaně (třídy) měly z celkového počtu konverzí více konverzí než ostatní.

2.3.2. Shlukování

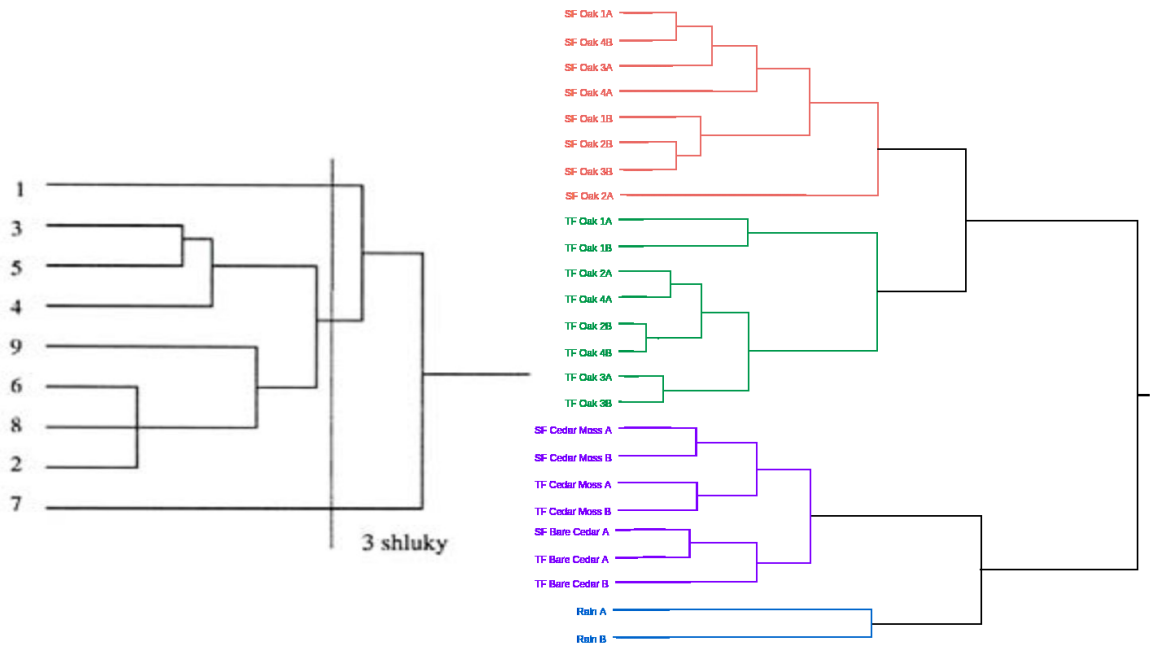
DM technika shlukování je typicky využívána pro úlohy segmentace. Cílem těchto úloh je rozdělit data do podskupin, které mají podobnou charakteristiku. Úlohy lze provádět manuálně na základě znalosti charakteru dat a odhadu jejich obchodního potenciálu, kdy tato znalost může být získána z dostupného popisu či jako výstup fáze deskripce a sumarizace dat, případně lze segmentaci provést s využitím technik automatického shlukování, které v datech detekují skryté struktury (Chapman et al., 2000, s. 67). P. Berka definuje shlukovou analýzu jako techniku, která hledá odpověď na otázku, zda a jak lze data rozdělit do shluků na základě vzájemně si blízkých příkladů. Vychází z předpokladu, že vzdálenosti mezi příklady, charakterizované numerickými veličinami, lze měřit. Vzdálenosti mezi dvěma příklady lze vyjádřit různými mírami (například lze využít Hammingovu vzdálenost, Eukleidovskou vzdálenost a další) (Berka, 2013, s.55-57).

Jednou z technik detekce skrytých struktur je shlukovací technika K-středy. Vzhledem k tomu, že algoritmus techniky neurčuje optimální počet shluků, je nutné počet shluků stanovit na počátku, tedy náhodně zvolit rozklad do *K* shluků. P. Berka uvádí, že z tohoto důvodu je technika méně náročná na výpočetní výkon než následně zmíněná technika hierarchického shlukování a tímto je vhodnější pro větší datové soubory. Algoritmus určí centroidy shluků, případně jsou centroidy určeny manuálně. Následně jsou jednotlivé příklady zařazovány do shluků dle jejich vzdáleností od určených centroidů. Variantou tohoto algoritmu je určit za centroidy prvních *K* příkladů či provádět přepočet centroidů po každém přesunu při zařazování příkladů do shluků dle jejich vzdáleností od centroidů. Shluky jsou následně reprezentovány svými centroidy, kdy tuto reprezentaci lze použít i pro zařazení nových příkladů. (Berka, 2013, s.59)

Další z technik detekce skrytých vzorů je hierarchické shlukování. P. Berka uvádí, že obvykle se při shlukování postupuje metodou “zdola nahoru”. Na začátku je každý příklad jedním shlukem. Tyto shluky na úrovni příkladů se postupně spojují až do dosažení jednoho velkého shluku obsahující všechny příklady. Jednotlivé shluky se k sobě spojují na základě nejbližších vzdáleností. Proces lze vizuálně znázornit pomocí tzv. dendogramu (viz Obrázky 7 a 8: Ukázky dendogramu), který zleva doprava znázorňuje postupné spojování shluků. Cyklus probíhá tak, že dokud existuje více než jeden shluk, algoritmus najde dva navzájem nejbližší shluky a ty spojí do jednoho. Pro tento nový shluk pak spočítá vzdálenost od ostatních shluků a takto pokračuje až do stavu, kdy zůstane jen jeden shluk.

Vzdálenosti lze určit následujícími způsoby. Jednou z metod je metoda nejbližšího souseda, kdy vzdálenost mezi shluky je dána minimem ze vzdálenosti mezi jejich příklady. Další metodou je metoda nejvzdálenějšího souseda, kdy vzdálenost mezi shluky udává maximum ze vzdálenosti mezi jejich příklady. Metodou průměrné vzdálenosti se vzdálenost mezi shluky určuje průměrem ze vzdálenosti mezi jejich příklady. U centroidní metody je vzdáleností mezi shluky vzdálenost mezi středy shluků. Centroidy fungují jako prototypy reprezentující jednotlivé shluky. Může se stát, že je jeden shluk reprezentován více centroidy (Berka, 2013, s.57-58).

Z obchodního hlediska může být typickou úlohou shlukování rozdělení zákazníků e-shopu do podskupin. Cílem je nalézt podskupiny, které jsou z obchodního hlediska zajímavé a mohou být užitečné pro další obchodní strategie, obzvlášť v současnosti, kdy již masové kampaně nejsou běžné a trend směřuje k demasifikaci reklamy. Pomocí shlukování může marketingové oddělení společnosti uspořádat data do zajímavých podskupin a tyto důkladněji analyzovat. Výstupy lze využít pro přesnější zacílení kampaní, pro optimalizaci umístění reklamy na webu za účelem zvýšení konverzního poměru či pro zvýšení povědomí o značce. Případně lze vytvořit pro samostatné skupiny specifické marketingové strategie - například lze na základě znalosti dat manuálně vytvořit segment e-mailových adres odebírající newsletter, jejichž uživatelé nakupují nejčastěji na území Prahy, kteří zároveň za poslední rok nakoupili alespoň ve stanoveném množství a jejich průměrná výše nákupu je větší, než stanovená, a odeslat jim voucher se slevou do nově otevřené pobočky v Praze. V případě, že je obchodním cílem napřímo oslovit tyto zákazníky v rámci otevření nové pobočky v Praze, výstup z této úlohy by byl zároveň i cílem celého *KDD* procesu. Slabým místem takové manuální aplikace může být subjektivita při určování kritérií výběru z celkové skupiny. Pouhý vstupní předpoklad, že skupina, která splňuje na vstupu zvolená pravidla, je jako cílová skupina nejrelevantnější, může zamezit objevení podstatných a na první pohled neviditelných skutečností a zkreslit výstup pro další obchodní strategie (Tai, 1997, s.105)



Obrázek 7: Ukázka dendogramu 1 (vlevo) (Berka, 2013, s.58)

Obrázek 8: Ukázka dendogramu 2: Dendrogram and Distance Cluster Analysis (vpravo) Dostupné online z: https://online.visual-paradigm.com/fr/diagrams/examples/dendrogram/dendrogram-and-distance-cluster-analysis/

U velké části zadání bývá technika shlukování jen jedním z kroků celého řešení, obzvlášť u úloh klasifikace či predikce. V tomto případě bývá cílem shlukování vytvořit z dat vhodné podskupiny pro další použití během klasifikace či predikce, což jsou úlohy řadící se k *discovery-driven data mining* technikám (Tai, 1997, s.105). Důvodem shlukování je pak nejen optimalizace řešení z pohledu výkonu, což jsou podobné důvody jako u metodologie *SEMMA*, kdy se v prvním kroku procesu k aplikaci vybírá pouze vzorek dat, ale především lze výběrem co nejrelevantnější podskupiny eliminovat možné zkreslení výstupů analýz nerozpoznáním zajímavých vzorů, ke kterému může vlivem vzájemného ovlivňování docházet, obzvlášť u velkých souborů dat. Segmentace je obzvlášť klíčová u úloh využívající asociační pravidla. Závislostní analýza na milionech záznamů je mnohem jednodušší, když se pro tyto účely vybere pouze homogenní podskupina dat omezená relevantní podmínkou, například časovým obdobím či hodnotou konverze. Výstupy z této analýzy jsou pak mnohem smysluplnější a přesnější (Chapman et al., 2000, s. 67).

2.3.3. Asociační pravidla

Asociační pravidla jsou společně s rozhodovacími stromy jednou z nejvyužívanějších data miningových technik. Technika pracuje s diskrétními hodnotami a případné numerické atributy se musí diskretizovat (Berka, 2003, s.102). P. Berka dále uvádí, že tento termín byl zpopularizován v souvislosti s algoritmem apriori, který navrhl R. Agrawal v souvislosti analýzou nákupního košíku. Algoritmus hledá, jaké druhy zboží si zákazníci kupují současně. Pomocí asociačních pravidel jsou hledány možné vazby mezi položkami, kdy žádná z položek není upřednostňována jako závěr pravidel. Levá strana reprezentuje předpoklad a pravá strana závěr. Generování kombinací (konjunkcí) hodnot, je základem všech algoritmů. Pravidlo lze vyjádřit čtyřpolní kontingenční tabulkou, kdy hodnoty v této tabulce vyjadřují frekvence kombinací. První hodnota vyjadřuje počet příkladů, kdy platí, že je splněn předpoklad i závěr, druhá hodnota vyjadřuje počet příkladů, kdy platí, že je splněn předpoklad, ale není splněn závěr, třetí hodnota vyjadřuje počet příkladů, kdy platí, že není splněn předpoklad, ale je splněn závěr a čtvrtá hodnota vyjadřuje počet příkladů, kdy není splněn ani předpoklad, ani závěr. Na základě těchto hodnot lze kvantitativně zhodnotit nalezené znalosti (Berka, 2003, s.102,103,106).

P. Berka uvádí, že při generování kombinací se prohledává prostor všech přípůstných kombinací a to buď metodou do šířky, do hloubky a nebo heuristickou. Při generování kombinací do šířky se generují kombinace dle délek, které začínají od délky jedna. Kategorie jednoho atributu jsou uspořádány podle abecedy. Při generování do hloubky se vychází od první kombinace délky jedna, která se prodlužuje o první kategorii dalšího atributu. Nelze-li kombinaci dále prodlužovat, mění se kategorie posledního atributu, případně se kombinace zkrátí a změní se hodnota kategorie posledního atributu u zkrácené kombinace. I v tomto případě jsou kategorie jednoho atributu uspořádány podle abecedy. Tyto způsoby generují kombinace na základě seznamu hodnot atributů, mohou tedy průběžně vznikat i kombinace s nulovou četností. Generování kombinací, kdy kombinace s nulovou četností jsou až na konci, tedy v pořadí dle četností v datech, je metodou heuristickou (Berka, 2003, s.106-107). Vzhledem k tomu, že generování kombinací je náročné na výpočetní výkon, lze počet kombinací redukovat požadavkem kombinace do určité délky či s určitou minimální četností. I přes redukování prohledávaného prostoru lze říct, že počet generovaných kombinací je exponenciálně závislý na počtu atributů (Berka, 2003, s.109). Výstupy z

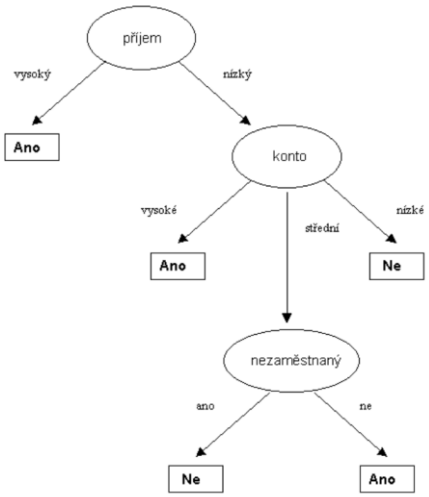
algoritmů detekující asociace většinou obsahují velké množství vazeb a výběr těch relevantních není vždy jednoduchý.

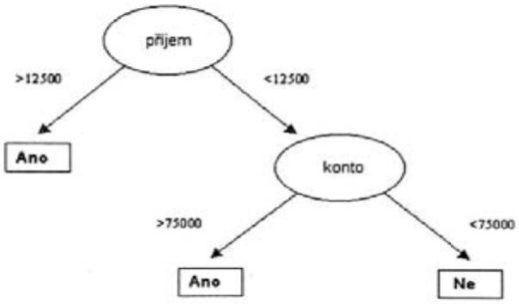
Příkladem využití asociačních pravidel v oblasti marketingu a e-commerce může být e-shop, který využívá asociační pravidla pro detekci skupin produktů, které zákazníci nejčastěji kupují dohromady. Tyto vazby pak využívají v rámci cross-sellingu, kdy pomocí e-mailů zákazníkům rozesílají doporučení k nákupu souvisejícího zboží. Případně lze tyto výstupy asociačních pravidel aplikovat přímo na webu e-shopu implementací odkazů u produktu na související zboží. Další forma využití může být vytvoření balíčků produktů, které mají mezi sebou vazbu, a nabízet je za výhodnější ceny.

2.3.4. Rozhodovací stromy

Rozhodovací stromy jsou vhodnou technikou pro úlohy klasifikace (Berka, 2003, s.130). Úlohy klasifikace pomocí klasifikačních modelů přiřazují veličinám jejich dříve neznámé diskrétní hodnoty. Při klasifikaci je předpokládáno, že existuje veličina, která obsahuje informaci o zařazení objektů do tříd - cílové veličiny. Ostatní veličiny jsou vstupními veličinami (Berka, 2003, s.66). Obecně úlohy klasifikace a techniky jejich řešení úzce souvisí s jinými typy data minigových úloh, které lze snadno převést na úlohy klasifikace (Chapman et al., 2000, s. 69).

P. Berka dále uvádí, že rozhodovací stromy fungují na principu rozděl a panuj. Obecný algoritmus rozhodovacích stromů pracuje s kategoriálními veličinami (viz Obrázek 9: Úplný rozhodovací strom pro kategoriální veličiny) a využity mohou být jen data bez šumu. U kategoriálních veličin může být veličina pro větvení stromu vybrána pouze jednou. U rozhodovacího stromu pro numerické veličiny (viz Obrázek 10: Úplný rozhodovací strom pro numerické veličiny) se veličiny v jedné větvi mohou opakovat. Trénovací data jsou rozdělena na menší podmnožiny, uzly stromu, tak, aby v každé podmnožině převládaly příklady jedné třídy. Tento postup je označován jako shora dolů a funguje pro kategoriální data, kdy cílem je z jednoho uzlu - kořene - rozdělit data na podmnožiny a vytvářet další uzly až do doby, než se v uzlu nachází pouze data jedné třídy. Přednost mají menší stromy (Berka, 2003, s.86).





Obrázek 9: Úplný rozhodovací strom pro kategoriální veličiny (vlevo) (Berka, 2013, s.93)
Obrázek 10: Úplný rozhodovací strom pro numerické veličiny (vpravo) (Berka, 2013, s.97)

Klíčové pro rozdělení dat do podmnožin je nalézt takový atribut, který od sebe jednotlivé příklady tříd nejlépe oddělí. Pro tyto účely lze využít například entropii, informační zisk či poměrný informační zisk. Entropie vyjadřuje míru neuspořádanosti nějakého systému a pro větvení stromu je vybrán atribut s nejmenší entropií. Informační zisk či poměrný informační zisk jsou míry odvozené z entropie a jedná se o rozdíl entropie pro cílový atribut a pro uvažovaný atribut. Pro větvení stromu je vybrán atribut s největší hodnotou informačního zisku. Poměrný informační zisk uvažuje i počet hodnot atributu. Vytvořený strom lze převést na sadu rozhodovacích pravidel. Nelistové uzly (atributy) jsou předpoklady pravidla a listové uzly (cíl) jsou závěry pravidla. Při dodržení výše uvedeného postupu větvení skončí až v momentu, kdy všechny příklady v listových uzlech patří do stejné třídy. Někdy však tento postup může vést k přeučení a strom může být nesrozumitelný. Řešením je prořezávání stromu a mít při implementaci v listovém uzlu příklady jedné třídy, které pouze "převažují". V důsledku je výsledný strom menší a srozumitelnější. Negativem je zhoršená klasifikace trénovacích dat. Prořezání může být dosaženo buď přímo vytvořením již redukovaného stromu či prořezáním uplného stromu. Klíčové je určit, kdy nelistový uzel nahradit listem. Toho lze docílit například prořezáváním pomocí algoritmu, který odhaduje vhodnost prořezávání na základě trénovacích dat, kdy odhaduje klasifikaci dat neznámých (Berka, 2003, s.87-95).

P. Berka uvádí, že přestože obecný algoritmus rozhodovacích stromů pracuje jen s kategoriálními veličinami, lze vytvořit rozhodovací strom i pro numerické veličiny. Zde vzniká problém vzhledem k velkému počtu možných hodnot, jelikož nelze pro každou hodnotu vytvořit samostatnou větev jako u kategoriálních veličin. Řešením je rozdělení hodnot na intervaly, v nejjednodušším případě na dva intervaly, které fungují jako diskrétní hodnoty. Zde je pak klíčové nalézt hodnotu dělícího bodu, který rozdělí hodnoty do intervalů. K nalezení tohoto bodu lze využít střední entropii atributu, kdy dělící bod má nejmenší (Berka, 2003, s.95-96).

Příkladem využití je marketingové oddělení e-shopu, které může pomocí rozhodovacích stromů z obecných informací o chování zákazníků na webu odvodit jejich předpokládaný věk, pohlaví, zájmy, oblíbené skupiny produktů či předpovědět, zda je uživatel potencionálním zákazníkem, tedy nakoupí, či ne. Pomocí rozhodovacích stromů lze vytvořit nástroj pro klasifikaci jak stávajících, tak i nových případů, a na tyto skupiny pak aplikovat různé obchodní strategie.

2.3.5. Regresní analýza

Úspěšnost výkonnostních kampaní měřená konverzním poměrem může být v korelaci s dalšími atributy, jako je čas a období spuštěné reklamy, kanál či zacílení reklamy. Na základě těchto veličin lze pomocí prediktivních metod odhadnout úspěšnost výkonnostních kampaní plánovaných v následujícícm období. Úlohy prediktivních analýz, na rozdíl od úloh klasifikace, jejichž výstupem jsou diskrétní veličiny, mají na výstupu veličiny spojité. Cílem je nalézt parametry závislosti a určit neznámé kvantitativní hodnoty zvolené veličiny. Tento typ úloh se nejčastěji řeší regresní analýzou dat. V případě, že součástí analýzy jsou i časové řady, je cílem výstup, který se označuje jako *forecasting* (Chapman et al., 2000, s. 70).

Regresní analýzy lze provádět pomocí lineární či nelineární regrese. P. Berka uvádí, že v případě lineární regrese jsou hledány parametry lineární závislosti mezi numerickými veličinami. Jedná se úlohu aproximace pozorovaných hodnot vyjádřenou funkcí s neznámými parametry. Lineární regrese pro dvě veličiny je nejjednodušším typem regrese. V případě lineární regrese jsou hledány hodnoty dvou parametrů, jejichž hodnotu lze

odhadnout, pokud jsou k dispozici vhodná pozorování dvojice hodnot. K těmto účelům lze využít metody nejmenších čtverců, která minimalizuje rozdíly mezi pozorovanou hodnotou a očekávanou hodnotou. U nelineární regrese je předpokládána složitější funkční závislost a to například kvadratická, obecně polynomická, exponenciální či logistická. Z pohledu *KDD* má větší význam mnohorozměrná regrese. U mnohorozměrné lineární regrese je předpokládáná lineární závislost závislé veličiny (cílového atributu) na nezávislých veličinách (vstupních atributech) (Berka, 2003, s.49-51).

Z podkapitoly vyplývá, že v ranných fázích procesu lze pro znázornění charakteristiky dat využít popisnou statistiku a vizualizační techniky, z jejichž výstupů lze definovat prvotní hypotézy o možných skrytých informacích v datech. Skryté informace v datech v oblasti marketingu lze získat pomocí několika data miningovými technik, například v podkapitole vybrané a uvedené techniky shlukování pro tvorbu segmentů, asociační pravidla, rozhodovací stromy pro klasifikaci či regresní analýzu pro predikci. U některých typů zadání bývají výstupy z jedné data miningové techniky využity jako vstup pro jiné. Například technika shlukování může být součástí úlohy klasifikace či výstupy segmentace mohou být použity v rámci aplikace asociačních pravidel. Volba vhodné modelovací techniky je v rámci procesu klíčová a přímo ovlivňuje relevantnost výstupu z obchodního hlediska. Vhodná volba je často závislá na znalostech a zkušenostech řešitele úlohy.

Data mining, z pohledu metodologie *CRISP-DM* fáze modelování, je součástí celého procesu dobývání znalostí z databáze a pro dosažení relevantních výstupů z obchodního hlediska ji nelze aplikovat odděleně. Techniky data miningu lze využít k různým typům úloh a obchodním cílům. Z pohledu marketingu nejčastěji za účelem personalizace reklamy a komunikace se zákazníky. Vzhledem ke komplexnosti úloh vzniklo na přelomu tisíciletí několik podpůrných metodologií *KDD* procesu, kdy nejvyužívanější je metodologie *CRISP-DM*, která je svým charakterem univerzální. Správná volba data miningové techniky má přímý vliv na splnění cílů z obchodního hlediska a někdy jeho dosažení vyžaduje několik iterací procesu.

3. Vybrané příklady využití DM v marketingu

Šest případových studií popisuje příklady aplikací v oblasti marketingu s využitím nejčastěji používaných data miningových technik a to rozhodovací stromy a shlukování.

Tyto techniky jsou využity i v rámci praktické části diplomové práce. Příklady jsou rozděleny dle primární data miningové techniky, kdy v některých případech jich bylo využito několik. Cílem všech aplikací je z obchodního hlediska využít data miningové techniky k personalizaci reklamy a udržování dlouhodobých vztahů se zákazníky, kterého, vzhledem k aktuálně velkému množství generovaných dat, nelze dosáhnout jinak, než s využitím informačních systémů a technik data miningu.

3.1. Příklady aplikace rozhodovacích stromů v marketingu

Případové studie v této podkapitole pracují s daty, která jsou měřena a ukládána pomocí *CRM* nástroje, což je nástroj, který shromažďuje informace o zákaznících jak osobního charakteru, tak i obchodního charakteru, za účelem tvorby obchodních a marketingových strategií společnosti dle aktuálních potřeb jednotlivých zákazníků. Cílem aplikace data miningu s využitím rozhodovacích stromů je v případových studiích klasifikace zákazníků na základě vzorů a charakteristik nalezených v datech.

3.1.1. Aplikace data miningu v přímém marketingu v bankovním sektoru

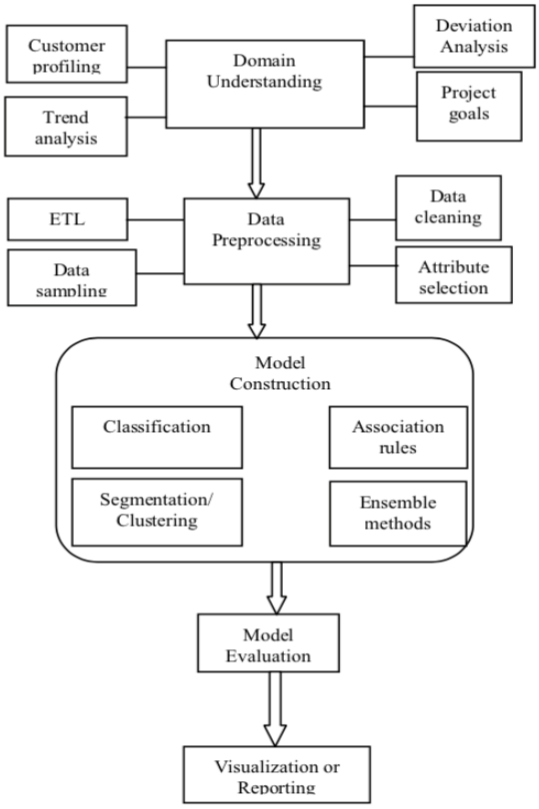
Jeden zdroj dat, data přímého marketingu portugalské banky, je použit ve třech dostupných případových studiích v rámci této sekce. Každá z případových studií má mírně odlišný cíl. První z nich je zaměřená na praktickou aplikaci a přináší konkrétní výstupy o nalezených skupinách, druhá z nich popisuje teoretický návrh procesu pro aplikaci data miningu v přímém marketingu, kdy praktická aplikace je popsána pouze okrajově, a ve třetí případové studii je cílem zvýšit pomocí data miningových technik účinnost marketingových kampaní a otestování dílčích výsledků dvou technik. Ve všech případech je použita jako jedna z data miningových technik rozhodovací stromy.

Úloha klasifikace má v první případové studii (Pavlovic et al, 2014) za úkol nalézt skupinu zákazníků, u kterých je velká pravděpodobnost, že do banky vloží dlouhodobý vklad a jsou ochotni účastnit se kampaní přímého marketingu zahrnující jejich oslovování a vyjádření názoru. Nalezená skupina zákazníků bude v rámci marketingových kampaní oslovena napřímo. Zadání úlohy je součástí obchodní strategie banky, která reflektuje trend budování dlouhodobých vztahů se zákazníky. Případová studie je řešena pomocí metod klasifikace a shlukování. Klasifikace je řešena data miningovou technikou rozhodovací

stromy a úloha shlukování technikou k-středů. Výstupy z úlohy shlukování jsou využity pro deskripci dat a zacílení kampaní na konkrétní skupiny klientů, zatímco výstupy z úlohy klasifikace jsou využity pro tvorbu predikcí, kdo z klientů provede dlouhodobý vklad. Slabou stránkou tohoto řešení mohou být potencionálně nepřesná vstupní data, která o sobě klient uvede.

Model rozhodovacího stromu je postaven na základě odpovědí klientů získaných v rámci kampaní přímého marketingu portugalské banky běžící za období květen 2008 - listopad 2010. Klienti banky byli osloveni telefonicky s nabídkou zajímavého úroku u dlouhodobého vkladu. V případové studii je využita jen část klientů, kteří mají o oslovování zájem a v minulosti již byli účastníky přímých marketingových kampaní. Na vstupu jsou jak osobní informace o klientovi, tak i informace marketingového charakteru o předchozí komunikaci s klientem. Klíčové veličiny jsou však pouze marketingového charakteru. Kritérii při konstrukci rozhodovacího stromu jsou délka posledního kontaktu, úspěšnost posledního kontaktu a měsíc, kdy ke kontaktu došlo. Výsledkem úlohy je zjištění, že klienti, co mají zájem o kontakt s bankovním agentem a délka kontaktu s ním byla delší, mají větší zájem o uzavření dlouhodobého vkladu. Klienti jsou dále metodou shlukování rozděleni do 12 skupin pomocí techniky K-středů a na tyto jednotlivé skupiny jsou aplikovány různé marketingové strategie reflektující charakter skupiny.

Ve druhé případové studii (Sing'oei, 2013) je popsán doporučený návrh procesu a technik pro data mining v oblasti přímého marketingu (viz Obrázek 11: Framework procesu pro aplikaci data miningu v oblasti přímého marketingu). Logika a obsah procesu odpovídá *KDD* procesu i metodologii *CRISP-DM*, liší se jen v označování. Praktická aplikace na bankovních datech je řešena okrajově a slouží jako příklad řešení. V praktické aplikaci je využit modelovací nástroj IBM SPSS. Původní dataset je rozdělen do deseti náhodně vybraných podmnožin, kdy devět podmnožin jsou tréninkové sady a jedna podmnožina testovací sada. Výstupem aplikace jsou dvě skupiny rozdělené na základě reakce s úspěšností zařazení přes 93%. Oproti první případové studii detailnější výstupy praktické aplikace nejsou k dispozici.



Obrázek 11: Framework procesu pro aplikaci data miningu v oblasti přímého marketingu (Sing'oei, 2013)

Ve třetí případové studii (Abbas, 2015), která také pracuje s datasetem přímého marketingu z oblasti bankovnictví, je hlavním cílem zvýšit účinnost marketingových kampaní a předvídat zájem klientů uzavřít dlouhodobý vklad. Pro optimálnější práci s daty, která jsou ve svém původním formátu uložena v textovém souboru, jsou data rozdělena do čtyř dílčích tabulek, která jsou mezi sebou propojena klíči. Ve dvou tabulkách jsou data osobního charakteru, ve třetí tabulce data bankovního charakteru a ve čtvrté tabulce data marketingového charakteru. Pomocí techniky teorie hrubých množin jsou stanoveny tři klíčové veličiny. Jedná se o veličiny s údaji o věku (data osobního charakteru), stavu účtu (data bankovního charakteru) a délce kontaktu (data marketingového charakteru). Pomocí těchto veličin jsou odvozena pravidla, která jsou po aplikaci hodnocena jako nejhodnotnější a s nejvyšší mírou přesnosti. Je zajímavé, že údaj o délce kontaktu byl vyhodnocen jako klíčový i v první případové studii. Dále je veličinám vypočtena jejich míra informační hodnoty. Veličina s největší mírou informační hodnoty je délka kontaktu, která je tak prvním uzlem v rozhodovacím stromu. Následně jsou odvozena další pravidla a jejich závěry na základě míry informační hodnoty jednotlivých veličin, avšak míra přesnosti jako u

kombinace klíčových veličin určených pomocí teorie hrubých množin dosažena nebyla. Ostatní veličiny hodnocené v teorii hrubých množin jako klíčové, věk a zůstatek na účtu, mají vypočtenou míru informační hodnoty až na osmém a desátém místě. Závěr studie je ten, že přestože je jednodušší implementovat konstrukci rozhodovacích stromů, výstupy nemají stejnou míru přesnosti jako při využití klíčových veličin odvozených pomocí teorie hrubých množin.

3.1.2. Data miningové techniky v oblasti Real-time marketingu

Případová studie (Gromov, 2011) popisuje návrhy řešení, které lze aplikovat v real-time marketingu z dat, která jsou získávána a ukládána pomocí *CRM* systému. Systém slouží k analýzám prodeje, nákupu, marketingu a dalších oblastí, ale neumožňuje nad daty provádět data miningové techniky a z výsledků předpovídat chování zákazníků jako podklad pro další marketingové strategie. Systém dále neumožňuje zpracovávat velké množství dat v reálném čase. Cílem případové studie je navrhnout řešení pro jednotlivé úlohy, které v současném systému nelze zpracovávat. První úlohou je vytvořit model, který za účelem zvýšení návratnosti investice odhaduje, koho oslovit přes kanál přímého marketingu. Druhou úlohou je získat odhad, které související produkty a služby zákazníci společně nakupují, jako podklad pro křížový prodej za účelem maximalizace prodeje. Třetí úlohou je segmentace a profilace zákazníků za účelem personalizace marketingu a tvorby strategií jednotlivých kampaní. Čtvrtou úlohou je odhad poptávky a ziskovost jednotlivých zákazníků za určité časové období. Výstupem je podklad pro plánování nákupu v návaznosti na předpokládanou poptávku, tvorbu obchodních strategií a identifikaci klíčových zákazníků.

Řešení je popsáno na příkladu klienta, který chce inzerovat reklamu na webu, která má dvě varianty - buď se zobrazí varianta reklamy s prodejem za nízké ceny u všeho zboží, nebo varianta reklamy s vysokými cenami. Varianta reklamy s vysokými cenami se zobrazí pouze zákazníkům, u kterých je největší pravděpodobnost, že nakoupí. Míra pravděpodobnosti je určena na základě historických dat. Klíčovými veličinami jsou údaje o předchozích nákupech, data osobního a marketingového charakteru. V prvním kroku jsou data nahrána do automodelu, který provede výběr data miningové techniky. V dalším kroku jsou data exportována a integrována do IBM InfoSphere Streams operátoru, což je nástroj pro distribuci výkonu mezi neomezený počet výpočetních uzlů. V této aplikaci je aplikován na data model SPSS Modeller, který vybraný model aplikuje na data a vygeneruje výstup. S

novými daty se model neustále aktualizuje a zpřesňuje. Architektura řešení umožňuje zpracovávat data v reálném čase.

V podkapitole jsou rozepsány čtyři případové studie, které využívají k řešení techniku rozhodovací stromy. Vzhledem k různým cílům a vstupům případových studií nelze jejich výsledky mezi sebou porovnat. V první případové studii byli nalezeni klienti mající větší zájem o nabízenou službu. Ve druhé případové studii je popsán doporučený návrh procesu a technik. Třetí případová studie se věnovala především optimalizaci volby data miningové techniky se závěrem, že přestože je jednodušší implementovat konstrukci rozhodovacích stromů, výstupy nemají stejnou míru přesnosti jako při využití k řešení teorii hrubých množin. Čtvrtá případová studie nabízí návrh řešení pro implementaci technik data miningu v reálném čase.

3.2. Příklady aplikace shlukování v marketingu

Jedna z případových studií aplikace shlukování v oblasti marketingu přináší nový pohled na přístup k data miningovým technikám v případě malých a středních podniků. Ve druhé případové skupině jsou zákazníci shlukováni na základě zájmů odvozených z navštívených stránek a členství ve skupinách na sociálních sítích.

3.2.1. Nový přístup k segmentaci zákazníků v malých a středních firmách

Případová studie (Raluca, 2012) přináší návrh na využití metod shlukování v malých a středních podnicích za účelem personalizace reklamy zákazníkům. Cílem je navrhnout řešení, které nevyžaduje velké náklady na specializovaný software a odborný personál, přesto je dostatečně účinné a snadno aplikovatelné. Řešení má nejpřesněji definovat a identifikovat cílové skupiny za účelem vytvoření marketingové strategie, která charakteristiku skupiny reflektuje. V některých případech je segmentace a klasifikace dle jedné či více charakteristik velice jednoduchá, avšak v některých případech jsou vyžadovány složité shlukové analýzy s využitím specializovaného softwaru. Získávaná data o zákaznících jsou osobního charakteru, marketingového charakteru a obchodního charakteru. Na základě společných charakteristik jsou zákazníci seskupováni do skupin. Na jednotlivé skupiny jsou následně aplikovány různé marketingové strategie. Na rozdíl od klasifikace zde

nejsou využita žádná tréninková data, které obsahují předdefinované třídy a příklady, které by naznačovaly možné vztahy mezi nimi.

Řešením pro malé a střední podniky je empirická aplikace shlukové analýzy bez použití specializovaných softwarů. Návrh řešení spočívá v rozřazení zákazníků pouze na základě tří nejdůležitějších kritérií, která jsou umístěna na osy trojrozměrného modelu. Do něj se zakreslí charakteristiky skutečného, či hypotetického klienta, který reprezentuje celou skupinu (dle mého názoru se jedná o tzv. "centroid", který pro danou oblast reprezentuje konkrétní klient). Základem procesu je Paretovo pravidlo, které uvádí, že zhruba 80% tržeb pochází od 20% zákazníků. Klíčoví klienti jsou graficky vyneseni do trojrozměrného modelu. Počet klientů by měl být z praktickýh důvodů relativně malý, aby provedení segmentace nevyžadovalo speciální nároky. U začínajících společností lze popsat ideální klienty pouze teoreticky. Je doporučeno mít u klíčových klientů co nejpřesnější data. U ostatních klientů stačí data přibližná. Pokud jsou dva či více klíčových klientů na grafickém znázornění příliš blízko u sebe, je z nich vytvořen jeden bod v trojrozměrném modelu, jehož souřadnice se vypočítají buď průměrem nebo váženým průměrem souřadnic jednotlivých klientů. Poté, co jsou do modelu zaneseni všichni klíčoví klienti, je trojrozměrný prostor kolem nich rozdělen do jednotlivých oblastí buď na základě obchodních zkušeností, či intuice. Pokud jsou někteří klienti důležitější, než jiní, mohou získat větší prostor kolem sebe. Následně jsou do modelu přiřazeni i ostatní zákazníci a dle umístění v modelu jim je určena skupina. Tento proces je iterační a zahrnuje operace na principu pokus-omyl. Je nutné ji provádět, kontrolovat a dynamicky upravovat v korelaci s výsledky na reálném trhu. Segmentaci s využitím tohoto postupu nelze automatizovat a platí v případě, kdy malý počet zákazníků generuje velké procento příjmů společnosti.

3.2.2. Analýza chování zákazníků využívající mobilní peněženku za účelem personalizovaného cílení reklamy

Případová studie (Alexandre et al., 2018) využívá metodologii *CRISP-DM.* V praktické části je využit programovací jazyk R. Cílem je získat z dat, které generují mobilní peněženky, informace o preferencích zákazníků za účelem zacílení personalizovaných marketingových kampaní uživatelům přímo do těchto aplikací. Osloveni reklamou by měli být pouze ti uživatelé, kteří by mohli mít o nabídku zájem. Segmentace uživatelů pro marketingové účely je provedena na základě zájmů uživatele a věkové skupiny.

Zpracovávaná data obsahují informace osobního charakteru, transakční data a data ze sociálních sítí. Data ze sociálních sítí, jako jsou navštívené stránky a členství ve skupinách, umožňují určení zájmů uživatelů a vytvářet na jejich základě segmenty. Na základě věku jsou uživatelé rozděleni do osmi věkových kategorií. Shlukování je provedeno pomocí techniky K-středů a technikou maximalizace očekávání. Maximalizace očekávání ovšem rozřadila zákazníky dle jejich zájmů do jednotlivých segmentů lépe a zdá se efektivněji využitelná pro cílení marketingových kampaní dle skupin uživatelů.

V podkapitole jsou popsány dvě řešení úloh shlukování, kdy první z nich nevyžaduje velké náklady na specializovaný software a odborný personál, přesto je dostatečně účinné a snadno aplikovatelné. Umožňuje relativně přesně definovat a identifikovat cílové skupiny za účelem vytvoření marketingové strategie. Úskalím může být skutečnost, že proces platí pouze v některých případech a je postaven na principu pokus-omyl. Ve druhé případové studii jsou porovnány dvě data miningové techniky pro úlohu shlukování, techniku K-středů a techniku maximalizace očekávání se závěrem, že technika maximalizace očekávání rozřadila uživatele dle jejich zájmů do segmentů mnohem lépe, než technika K-středů.

Výstupy z šesti případových studií popsaných v kapitole přímo ovlivnily výběr technik pro úlohy segmentace a klasifikace řešené v rámci praktické části diplomové práce, která má podobné obchodní cíle, jako byly uvedené v případových studiích.

4. Praktická aplikace vybraných metod DM na marketingových datech dle metodologie CRISP-DM

Kapitola praktické aplikace pro kterou jsou využita vzorová data Google Analytics, která obsahují webová data a transakční data e-shopů měřená pomocí Google Merchandise Store, má za cíl nalézt skupinu uživatelů, pokud existuje, která reaguje na marketingovou komunikaci častěji, než ostatní. Praktická aplikace je řešena posloupností fází a kroků ve shodě s metodologií *CRISP-DM*. Ve fázi porozumění problematice jsou stanoveny cíle a přínosy a vypracován projektový plán. Ve fázi porozumění datům jsou nalezeny prvotní hypotézy, o kterou se opírá fáze modelování i interpretace. Fáze příprava dat upravuje data do formy vhodné pro fázi modelování. Ve fázi modelování jsou vybrány a aplikovány data miningové techniky za účelem splnění cílů stanovené ve fázi porozumění problematice.

Následujícím fázím zhodnocení a využití je věnovaná kapitola následující. Podkapitoly jsou rozděleny na jednotlivé sekce, které odpovídají jednotlivým krokům fází.

4.1. Porozumění problematice

V podkapitole jsou popsány cíle a přínosy praktické aplikace z obchodního hlediska a z hlediska data miningu, posouzen výchozí stav situace, popsány nástroje využité k řešení a nastíněn projektový plán dalších fází.

4.1.1. Stanovení cílů z obchodního hlediska

Hlavním cílem praktické aplikace z obchodního hlediska je nalézt, pokud existuje, skupinu uživatelů, kteří reagují na marketingové kampaně častěji než ostatní, nalézt případné souvislosti v databázi, které mají na tuto skutečnost vliv, a na základě těchto zjištění optimalizovat marketingovou strategii za účelem zvýšení konverzního poměru. V případě, že taková skupina nalezena nebude, lze zvýšit konverzní poměr popisem společných charakteristik uživatelů e-shopu a těmto skupinám přizpůsobit strategie plánovaných kampaní. Vedlejším obchodním cílem je odhadnout, zda nově akvirovaný uživatel provede transakci, či ne, a tomu již od počátku přizpůsobit komunikaci.

Přínosem praktické aplikace je popis zjevných a případné získání skrytých informací o návštěvnících na podkladě transakčních dat a dat webové analytiky, ze kterých mohou těžit marketingoví specialisté při plánování kampaní. Úspěšným výsledkem z obchodního hlediska je zvýšení konverzního poměru.

4.1.2. Posouzení situace

Zdrojem dat jsou data Google Analytics sample volně dostupná přes platformu Google BigQuery, která kombinují data webové analytiky s transakčními daty e-shopů. Získávání dat webové analytiky měřené v Google Analytics přes platformu Google BigQuery umožňuje získat nevzorkovaná data v kombinaci s daty třetích stran, například s transakčními daty e-shopů. Vzhledem k tomu, že se jedná o transakční data a data webové analytiky e-shopů z celého světa, je pro účely diplomové práce předpokládáno, že se jedná o jeden konzistentní e-shop s celosvětovým působením, který si aktuálně jako

marketingovou strategii zvolil masovou mediální kampaň s plošným zacílením na stávající i potencionální zákazníky přes veškeré mediální kanály (direct - přímý prodej, paid search - placené vyhledávání, social - sociální sítě, affiliate - provizní program, display - banerová reklama). Analýza nákladů a přínosů praktické aplikace není součástí této diplomové práce, byť v metodologii *CRISP-DM* je tato analýza nedílnou součástí tohoto kroku.

4.1.3. Stanovení cílů data miningu

Hlavní cíl praktické aplikace, nalezení skupiny uživatelů, kteří reagují na marketingové kampaně častěji, než ostatní, bude řešen pomocí úlohy segmentace. Segmentací jsou návštěvníci, kteří jsou v datasetu identifikováni dimenzí fullVisitorId, rozděleni do několika skupin a to na základě počtu návštěv uživatele, celkového počtu navštívených stránek, průměrného počtu navštívených stránek uživatelem během jedné návštěvy, celkové doby uživatele na webu, průměrně strávené doby uživatele na webu během jedné návštěvy a počtu transakcí. Popis nalezených skupin a případná identifikace jedné, která reaguje na marketingové kampaně častěji, než ostatní, je splněním hlavního cíle z pohledu data miningu. Dílčí výstupy agregací dat jsou splněním vedlejšího cíle z hlediska data miningu a to určení obecné typologie uživatelů, kteří e-shop navštěvují.

Pro vytvoření odhadu, zda nový návštěvník provede transakci, je využito metody klasifikace pomocí techniky rozhodovacích stromů. Na základě trénovacích dat jsou odvozena pravidla rozhodovacího stromu. Výstupem jsou pravidla, která umožní odhad, zda nový návštěvník, označen v primárních datech hodnotou v dimenzi fullVisitorId, je uživatelem, který provede transakci. Splněním cíle z data miningového hlediska je přesnost modelu alespoň 70%.

4.1.4. Vytvoření projektového plánu

Pro fázi porozumění datům, za účelem analýzy dat a volby relevantních metrik a dimenzí, je stažen vzorek do lokálního zařízení z cloudové platformy BigQuery ve formátu JSON za období jednoho dne, který je rozparsován a prvotně zanalyzován pomocí programovacího jazyka R. Z tohoto vzorku je proveden pouze omezený výběr dimenzí a metrik relevantních k úloze. Vzorek obsahující relevantní dimenze a metriky je prozkoumán a interpretován pomocí vizualizačního nástroje Tableau pro vytvoření jednoduchých

agregací, nalezení trendů a možných souvislostí. Pomocí technik agregace dat a vizualizace v rámci fáze porozumění datům jsou popsány obecné charakteristiky zákazníků a jejich chování na webu. Získání dat z cloudovém prostředí Google BigQuery přes API a příprava dat na podkladě výstupů z fáze porozumění datům pro fázi modelování, je provedena dotazy SQL s využitím programovacího jazyka R. Fáze modelování je řešena s využitím platformy RapidMiner. Hledání skupiny uživatelů, kteří reagují na marketingovou komunikaci častěji, než ostatní, je řešeno pomocí segmentace. Pro vytvoření odhadu, zda nově akvirovaný návštěvník provede transakci, či ne, je využito metody klasifikace. Fáze využití je popsána pouze teoreticky, jelikož s nasazením do ostrého provozu a následným praktickým využitím se v rámci této diplomové práce nepočítá. Kritéria pro hodnocení úspěšnosti jsou splnění cílů z pohledu data miningu a splnění cílů z pohledu možného obchodního využití.

Z podkapitoly vyplývá hlavní cíl praktické aplikace, a to nalezení skupiny uživatelů, kteří reagují na marketingové kampaně častěji, než ostatní, a vedlejší cíl praktické aplikace, odhadnout, zda nový uživatel provede transakci, či ne. Splněním cíle z data miningového hlediska je nalézt segment, který odpovídá charakteristice hledané skupiny a u klasifikačního modelu přesnost alespoň 70%. Zdrojem dat jsou data Google Analytics sample volně dostupná přes platformu Google BigQuery získaná přes API s využitím programovacího jazyka R. V projektovém plánu jsou shrnuty následující kroky řešení.

4.2. Porozumění datům

V podkapitole jsou popsány použité zdroje dat a data, která obsahují. Jsou stažena iniciační data a vybrány z nich pouze relevantní dimenze a metriky pro splnění cílů z obchodního i data miningového hlediska. Následně jsou vytvořeny vizualizace za účelem prvotní charakterstiky dat a vytvoření hypotéz. Součástí této fáze je i ověření a zhodnocení kvality dat.

4.2.1. Shromáždění iniciačních dat

Pro prvotní porozumění datům jsou data za období jednoho dne stažena z cloudové platformy Google BigQuery přes API s využitím programovacího jazyka R a uložena do lokálního zařízení. Z technického hlediska se jedná se o jediný zdroj dat, avšak data, která tento zdroj obsahuje, jsou kombinací více zdrojů a to Google Analytics 360 a Google

Merchandise Store. Formát stažených dat do lokálního zařízení je JSON, který obsahuje vnořená pole a objekty. Pro účely porozumění datům jsou s využitím programovacího jazyka R data rozparsována a provedena analýza za účelem výběru relevantních dimenzí a metrik potřebných pro splnění obchodních i data miningových cílů, jejichž výčet je uveden ve fázi porozumění problematice. Výsledný upravený dataset, obsahující relevantní dimenze a metriky za období jednoho dne, je pro případné další použití lokálně uložen ve formátu csv. V tomto kroku jsou následně stažena vybraná iniciačních data za kompletní období jednoho roku.

4.2.2. Popis dat

Zdrojový dataset po rozparsování obsahuje přes 300 veličin s dimenzemi a metrikami zaznamenávající informace o zákaznících, návštěvnosti webu, jednotlivých stránkách, zdrojích návštěvy, cestě zákazníka na webu, jednotlivých transakcích, ceně objednávek, ceně jednotlivých produktů, informace o zařízeních, prohlížeči, operačním systému, geolokačních datech, metriky pro vyhodnocování kampaňových dat, časové údaje a další veličiny za období jednoho roku od srpna 2016 do srpna 2017. Dataset Google Analytics sample obsahuje kromě základních dimenzí a metrik webové analytiky a transakčních dat, jako je identifikátor návštěvníka, návštěvnost, navštívené stránky či počet transakcí, také přístup k vlastním dimenzím a metrikám, které lze v Google Analytics pro vlastní měření nastavit. Detailní rozbor veškerých dimenzí a metrik, který tento dataset obsahuje, není předmětem praktické části diplomové práce. V rámci praktické části jsou popsány pouze vybrané dimenze a metriky relevantní pro splnění cílů z obchodního a data miningového hlediska.

Dataset s relevantními dimenzemi a metrikami obsahuje dimenzi FullVisitorId, v datasetu pod označením Navstevnik, což je identifikátor návštěvníka, dimenzi visitId, v datasetu pod označením Navsteva, která je použita pro určení počtu návštěv zákazníka, dimenzi Date, v datasetu pod označením Datum, ze které je odvozena dimenze měsíc-rok, dimenzi channelGrouping, v datasetu pod označením ZdrojNavstevy, kde je uveden kanál, ze kterého zákazník na web přišel. Informaci o zdroji návštěvy lze získat i z jiných dimenzí, například Source (zdroj) a Medium, avšak pro splnění cílů postačí dimenze channelGrouping. Dataset dále obsahuje dimenzi device.deviceCategory, v datasetu pod označením Zarizeni. Dimenze hits.contentGroup.previousContentGroup2 je v datasetu pod

označením KategorieProduktu a obsahuje agregované údaje o obsahu, který uživatel navštívil na předchozí stránce. Z této dimenze lze odvodit na obecné úrovni zájem uživatele o jednotlivé kategorie produktů na webu. Obsah na předchozí stránce byl vybrán z důvodu, že je nutné přiřadit kategorii produktu u uživatelů, kteří provedli transakci. Stránka, která potvrzuje transakci údaje o kategorii produktu již neobsahuje. ID transakce je vytvořeno a přiřazeno na stránce, která potvrzuje objednávku. Absence údajů o obsahu poslední navštívené stránky u uživatelů, kteří transakci neprovedli, je vzhledem ke stanoveným cílům zanedbatelná. Dimenze hits.transaction.transactionId, v datasetu pod označením TransactionID, identifikuje provedené objednávky uživatelem. Metrikami v datasetu jsou údaje o době návštěvy uživatele na webu, o počtu navštívených stránek a údaje o příjmech za každou transakci (*Revenue*). Zmíněné veličiny jsou postačující pro vytvoření odvozených metrik potřebných v pozdějších fázích, aby byly splněny cíle z obchodního i data miningového hlediska. Z celého objemu vzorových dat, po výběru pouze relevantních dimenzí a metrik a jejich stažení přes API cíleným dotazem, dataset obsahuje necelých 1,4 milionů řádků a 11 sloupců (viz Obrázek 12: Dataset obsahující relevantní dimenze a metriky načtené z Google BigQuery přes API do prostředí R studia) zahrnující údaje o aktivitách necelých 360 tisících uživatelů, kteří během jednoho roku navštívili stránky e-shopu v rámci zhruba 460 tisíc návštěv a provedli necelých 12 tisíc transakcí.

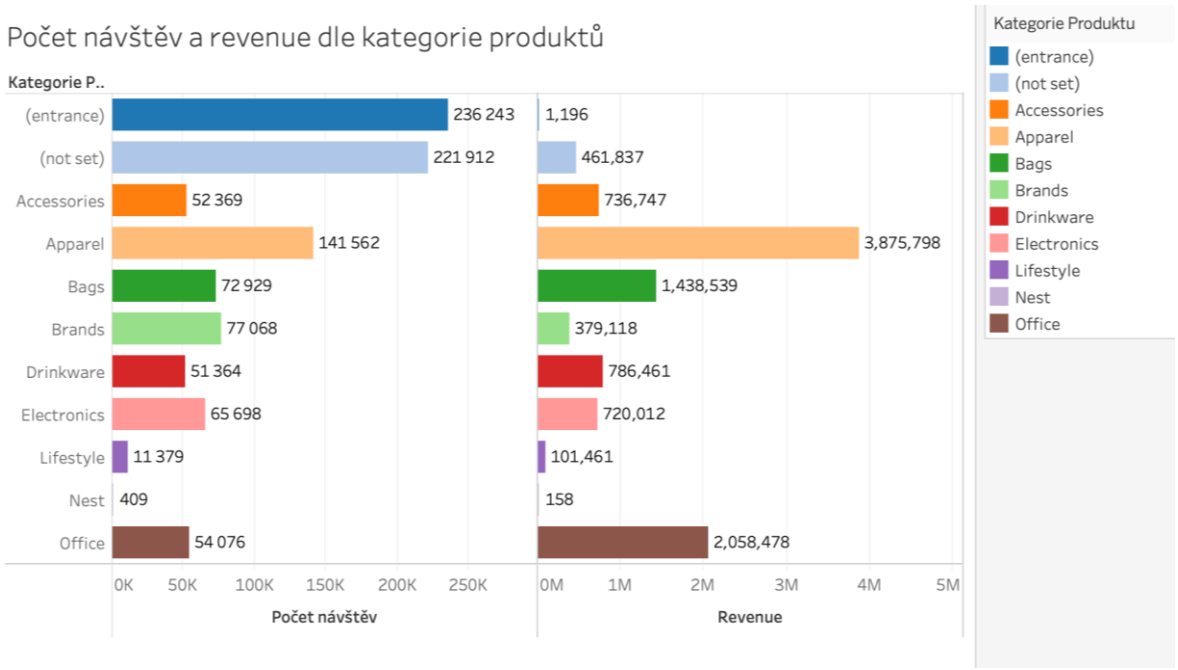
Navstevnik	Navsteva	Datum	ZdrojNavstevy	Zarizeni	KategorieProduktu	TypAkce	DobaNavstevy	PocetNavstivenychStranek	TransactionID	Revenue
0000010278554503158	1477029466	20161020	Organic Search	desktop	Accessories	2	194	8	NA	NA
0000010278554503158	1477029466	20161020	Organic Search	desktop	Accessories	1	194	8	NA	NA
0000010278554503158	1477029466	20161020	Organic Search	desktop	(not set)	0	194	8	NA	NA
0000010278554503158	1477029466	20161020	Organic Search	desktop	Accessories	0	194	8	NA	NA
0000010278554503158	1477029466	20161020	Organic Search	desktop	Electronics	0	194	8	NA	NA
0000010278554503158	1477029466	20161020	Organic Search	desktop	Office	0	194	8	NA	NA
0000020424342248747	1480578901	20161130	Organic Search	desktop	Bags	2	297	13	NA	NA
0000020424342248747	1480578901	20161130	Organic Search	desktop	Electronics	2	297	13	NA	NA
0000020424342248747	1480578901	20161130	Organic Search	desktop	Bags	1	297	13	NA	NA
0000020424342248747	1480578901	20161130	Organic Search	desktop	Electronics	1	297	13	NA	NA
0000020424342248747	1480578901	20161130	Organic Search	desktop	(not set)	0	297	13	NA	NA
0000020424342248747	1480578901	20161130	Organic Search	desktop	Apparel	0	297	13	NA	NA
0000020424342248747	1480578901	20161130	Organic Search	desktop	Bags	0	297	13	NA	NA
0000020424342248747	1480578901	20161130	Organic Search	desktop	Brands	0	297	13	NA	NA
0000020424342248747	1480578901	20161130	Organic Search	desktop	Electronics	0	297	13	NA	NA
0000027376579751715	1486866293	20170211	Organic Search	desktop	Accessories	2	49	5	NA	NA
0000027376579751715	1486866293	20170211	Organic Search	desktop	Accessories	1	49	5	NA	NA

Obrázek 12: Dataset obsahující relevantní dimenze a metriky načtené z Google BigQuery přes API do prostředí R studia

4.2.3. Prozkoumání dat

Pomocí vizualizací provedených ve vizualizačním nástroji Tableau jsou zjištěny prvotní informace, které lze z dat vyčíst. Z prvotních datových charakteristik jsou odvozeny

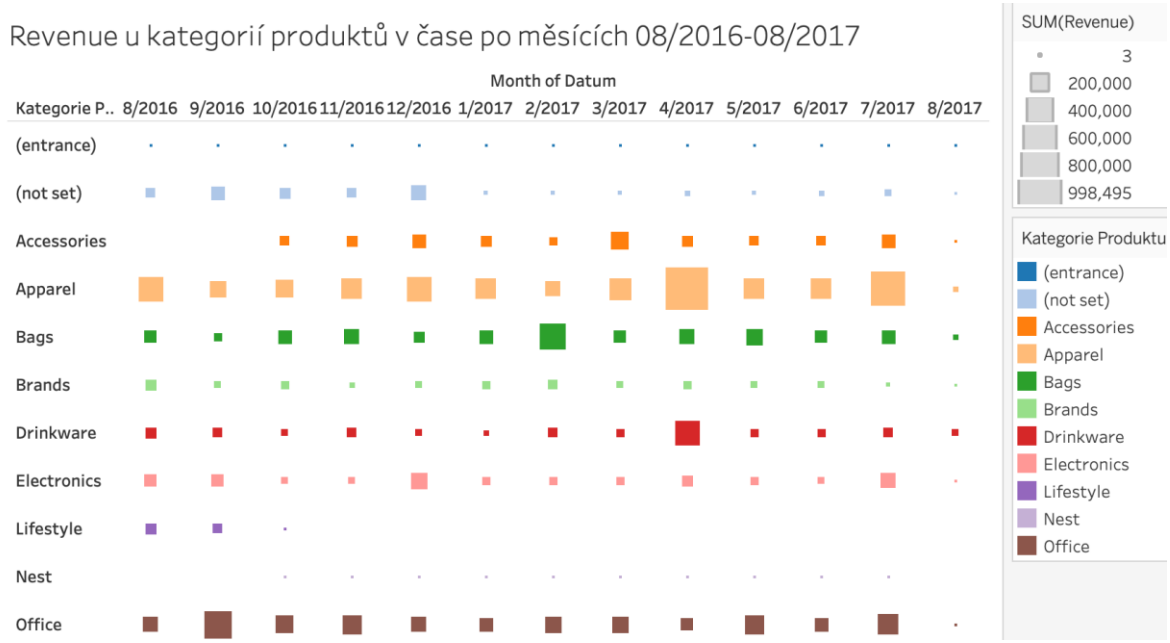
počáteční hypotézy o chování uživatelů na webu. Z vizualizace na Obrázku 13 (Počet návštěv a Revenue dle kategorie produktů) vyplývá, že návštěvníci webu nejvíce navštěvují stránky s oblečením, na druhém místě se umístily brandové stránky a na třetím místě jsou kabelky a tašky. Skutečný zájem o kategorie produktů se promítají v Revenue, tedy v příjmech za jednotlivé kategorie. Největší příjem má kategorie oblečení, na druhém místě jsou kancelářské potřeby a na třetím místě jsou kabelky a tašky. Důvod, proč Revenue v kategorii brand poměrově neodpovídá návštěvnosti, jako je tomu třeba u oblečení či kabelek a tašek je ten, že primárním cílem brandových stránek není prodej, ale rozšíření povědomí o značce. Z těchto zjištění lze určit hypotézu, že skupina uživatelů, která reaguje na marketingovou komunikaci častěji, než ostatní, má mezi zájmy s určitou pravděpodobností oblečení, kabelky a tašky.



Obrázek 13: Počet návštěv a Revenue dle kategorie produktů

Další zajímavé charakteristiky dat lze vyčíst z grafu na Obrázku 14 (Revenue u kategorií produktů v čase po měsících) znázorňující vývoj Revenue u kategorií produktů po měsících. U nejoblíbenější kategorie oblečení je Revenue ve většině případů ze všech kategorií nejvyšší, pouze ve dvou případech je druhým v pořadí, poprvé v září 2016, kdy tuto kategorii předstihla kategorie kancelářských potřeb (možná z důvodu korelace se začátkem školního roku) a podruhé v únoru 2017, kdy kategorii oblečení předstihla kategorie

kabelek a tašek. Nejvyšší hodnotu mělo Revenue v dubnu 2017 kategorie oblečení. Je zajímavým jevem, že v dubnu 2017 oproti jiným měsícům vykazovala výrazně vyšší Revenue kategorie sklenic a hrnků.

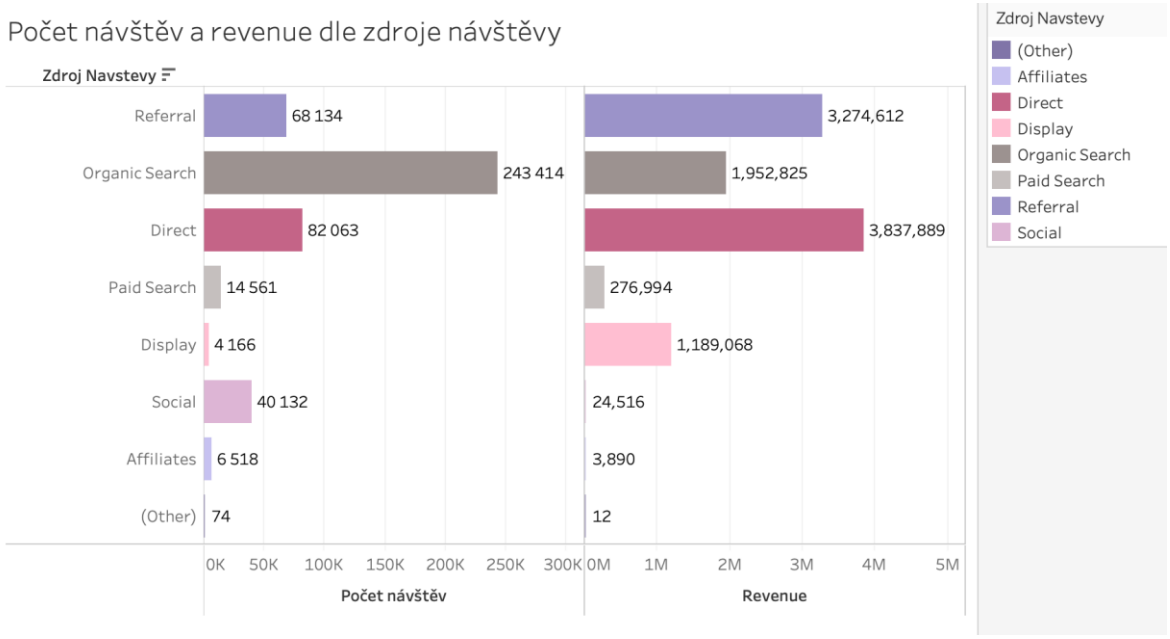


Obrázek 14: Revenue u kategorií produktů v čase po měsících

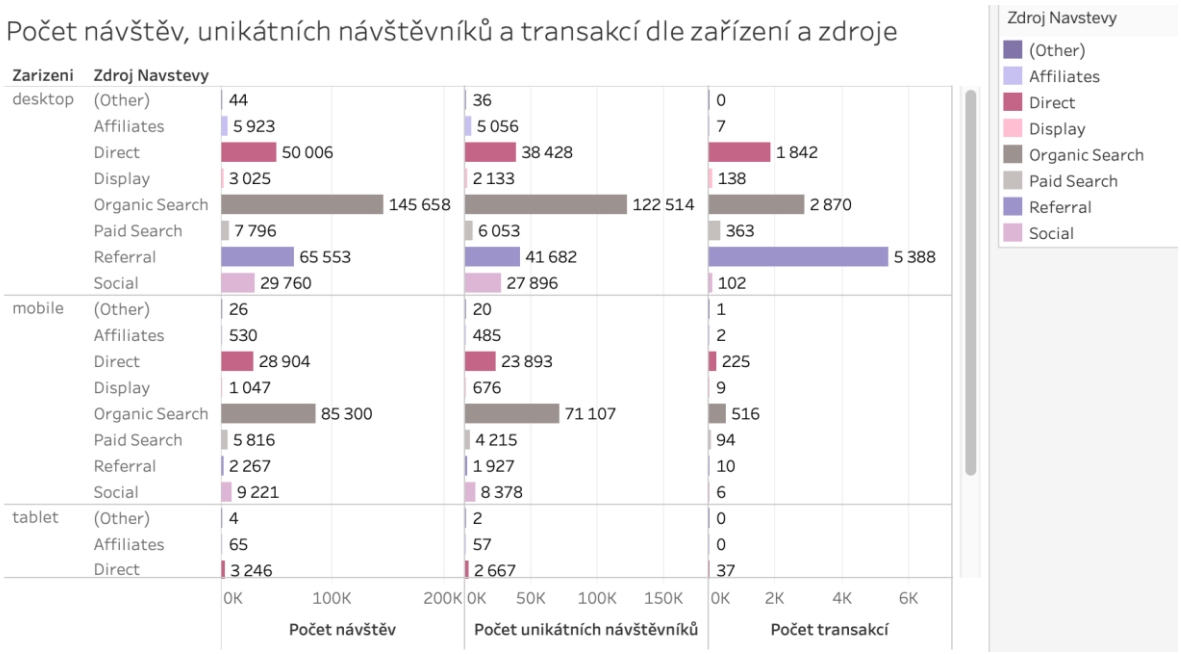
Vizualizace dat na Obrázku 15 (Počet návštěv a Revenue dle zdroje návštěvy) potvrzuje úspěšnost aktuálního trendu péče o dlouhodobé zákazníky. Přestože počet návštěv z kanálu Direct je až třetí v pořadí, tento kanál má nejvyšší Revenue ze všech, zřejmě v důsledku kampaní přímého marketingu komunikující napřímo s loajálními zákazníky. Z vizualizace dat lze dále vyčíst, že sice je většina návštěv nekampaňová, příjmy z těchto návštěv jsou až třetí v pořadí. Na základě těchto charakteristik lze vytvořit hypotézu, že skupina uživatelů, kteří reagují na marketingovou komunikaci častěji, než ostatní, jsou zákazníci, se kterými je komunikováno pomocí přímého marketingu, např. e-mailem.

Z vizualizace dat na Obrázku 16 (Počet návštěv, unikátních návštěvníků a transakcí dle zařízení a zdroje) vyplývá, že přestože nejvíc uživatelů přijde na web na základě organického vyhledávání ať už z mobilu, či desktopu, nejvíce transakcí na desktopu proběhne z marketingového kanálu *Referral* (referenční odkazy směřující na web e-shopu umístěné na webu různého charakteru, ať už blogy, sociální sítě a další). Na mobilním zařízení si poměr počtu návštěv a počtu transakcí napříč kanály přibližně odpovídá. Nejsilnějším kanálem na mobilu je *Organic search* (organické vyhledávání). Z grafu dále

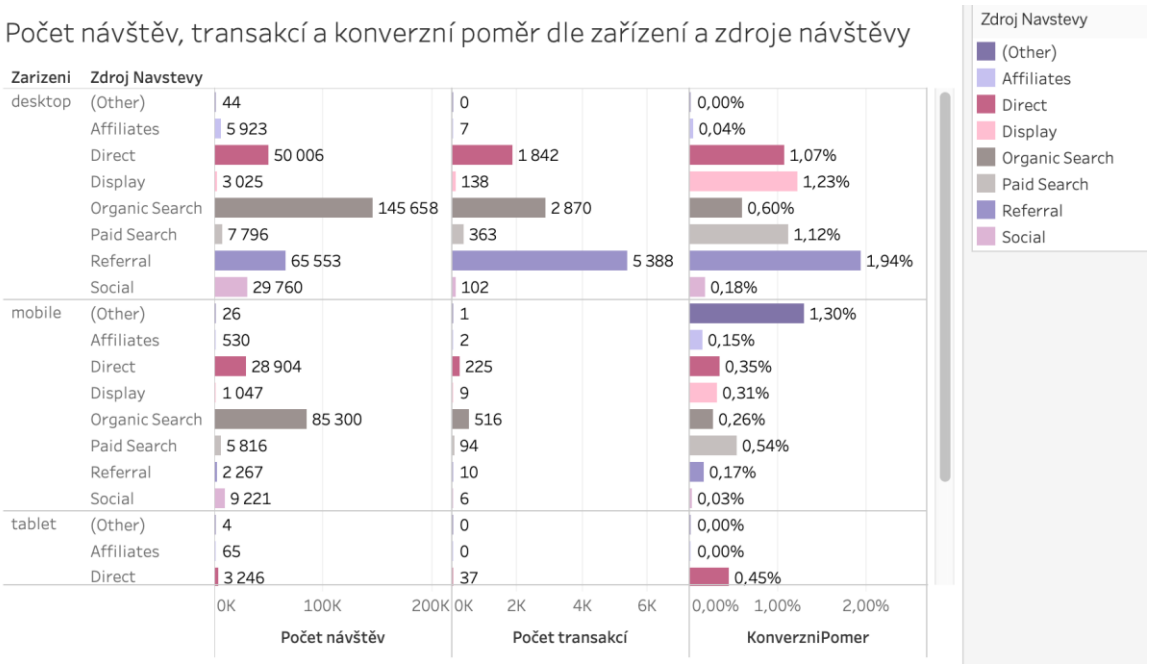
vyplývá, že uživatelé desktopu, se kterými je komunikováno pomocí kanálu přímého marketingu, web navštíví v průměru více než jednou. Na desktopu má kanál Referral největší počet transakcí jak napříč kanály, tak i zařízeními. Vizualizace na Obrázku 18 (Počet transakcí a Revenue dle zařízení a zdroje návštěvy) znázorňuje, že byť má kanál přímého marketingu nejvyšší Revenue, je počtem transakcí až třetí v pořadí. Zde je hypotéza, že důvodem je to, že vhodně zacílené kampaně přímého marketingu motivují uživatele k transakcím s vyšší hodnotou.



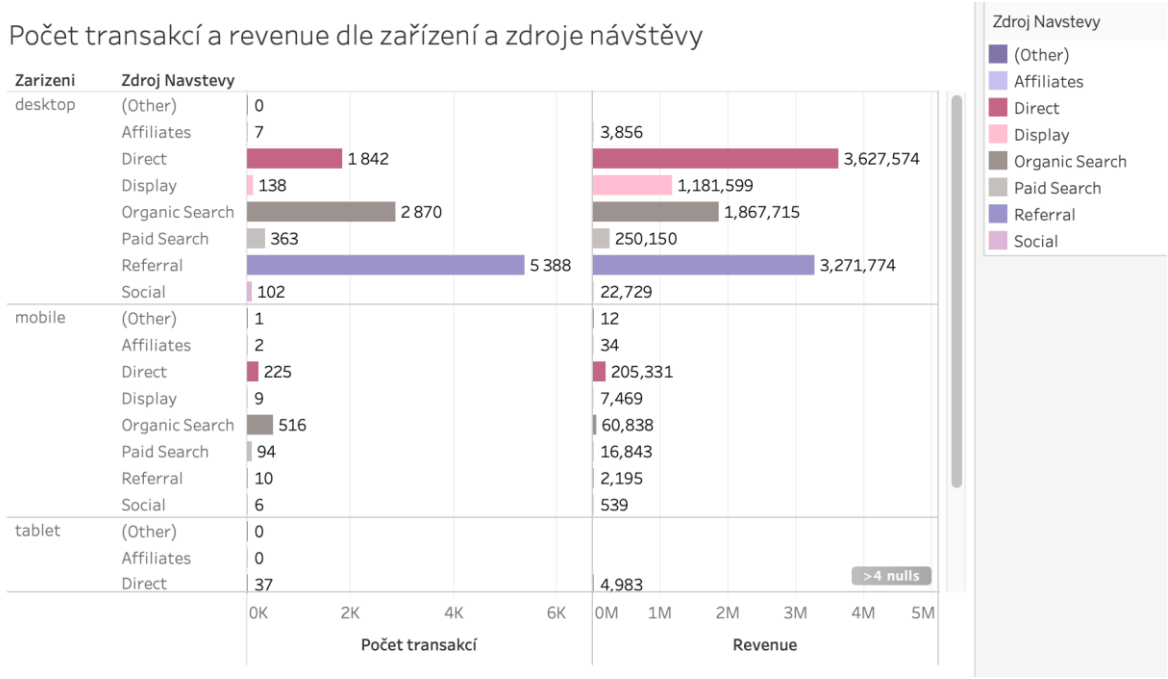
Obrázek 15: Počet návštěv a revenue dle zdroje návštěvy



Obrázek 16: Počet návštěv, unikátních návštěvníků a transakcí dle zařízení a zdroje



Obrázek 17: Počet návštěv, počet transakcí a konverzní poměr dle zařízení a zdroje návštěvy



Obrázek 18: Počet transakcí a Revenue dle zařízení a zdroje návštěvy

Na základě vizualizace na Obrázku 17 (Počet návštěv, počet transakcí a konverzní poměr dle zařízení a zdroje návštěvy) lze vytvořit další hypotézu, že skupina uživatelů, co reaguje na marketingové kampaně častěji, než ostatní, využívá kromě kanálu přímého marketingu i kanál Referral na desktopovém zařízení. Hypotéza vznikla na základě toho, že tento kanál má na desktopovém zařízení výrazně vyšší konverzní poměr oproti ostatním.

Ze souhrnů dat (v Příloze 2) lze vyčíst, že u nejoblíbenější kategorie oblečení je medián Revenue za jednu transakci mezi kategoriemi produktů druhý nejnižší, avšak kategorie má nejvyšší průměrnou četnost transakcí za měsíc. Druhá nejoblíbenější kategorie kancelářské potřeby má oproti oblečení třetinovou průměrnou četnost, ale zároveň má medián revenue za jednu transakci jeden z nejvyšších.

Na základě vizualizace a popisné statistiky za účelem určení prvotních charakteristik dat je vytvořeno několik prvotních hypotéz, že uživatelé, kteří reagují častěji, než ostatní na marketingovou komunikaci, tedy vytvořili alespoň jednu transakci, jsou uživateli se zájmem o oblečení, kabelky, tašky a kancelářské potřeby, se kterými je komunikováno pomocí přímého marketingu či na web chodí přes marketingové referenční odkazy. Tito uživatelé nakupují kontinuálně v průběhu celého roku, v průměru několikrát za měsíc, mají vysoký konverzní poměr, jejich jednotlivé objednávky nemají vysokou hodnotu a nejčastěji k takovým nákupům využívají desktopové zařízení. Na webu stráví přibližně 20 min, během kterých navštíví kolem 35 stránek. Zařazením těchto uživatelů do samostatné skupiny umožní připravit a zacílit speciální kampaně přímého marketingu, které tyto charakteristiky skupiny budou reflektovat.

4.2.4. Ověření kvality dat

Iniciační data obsahují pouze relevantní dimenze a metriky. Vyskytující se prázdné hodnoty u dimenze TransactionID a metriky Revenue znamenají absence transakce a neznamenají chybějící hodnoty v pravém slova smyslu. Dimenze KategorieProduktu obsahuje kromě kategorií produktů také hodnoty “(not set)” a “(entrance)”. Hodnota “(not set)” znamená, že k danému produktu není přiřazena kategorie. Hodnoty “(entrance)” jsou důsledkem použití dimenze contentGroup předchozí stránky, kdy důvodem použití této dimenze je absence údajů o kategorii na stránce při provedení transakce, tzn. za účelem přiřazení kategorie transakci. Hodnoty “(entrance)” se přiřazují uživatelům, co vstoupili na

stránku, následně šli na další stránku a poté web opustili. Ve většině případů jde o uživatele bez provedené transakce. Pouze ve čtyřech případech u této kategorie k transakci došlo a zde mě napadá jako důvod například ten, že uživatel přes vstupní stránku pokračoval rovnou do košíku, kde měl uložen výběr z předchozí návštěvy.

Z podkapitoly vyplývá, že jsou z Google BigQuery přes API stažena, prozkoumána a vybrána iniciační data, jejichž kvalita je dostatečná pro splnění obchodních i data miningových cílů. Z iniciačních dat jsou vytvořeny prvotní vizualizace a souhrny, ze kterých jsou zjištěny zajímavé hypotézy o skupině uživatelů, kteří by mohli reagovat na marketingovou komunikaci častěji, než ostatní.

4.3. Příprava dat

V rámci fáze přípravy dat je popsána příprava dat a datasetů obsahující pouze numerické hodnoty, který jsou použity ve fázi modelování. V této fázi je dále připraven rozšířený dataset pro fázi zhodnocení výsledků a popsány důvody, proč není rozšířený dataset použitý již ve fázi modelování.

4.3.1. Výběr dat

Metoda pro úlohu segmentace technikou automatizovaného shlukování pomocí K-středů je omezená na nekategoriální data, tzn. numerická data. Z tohoto důvodu je nutné tomuto omezení dataset přizpůsobit. Dataset na vstupu má 11 veličin, jejichž hodnoty jsou násobeny až do úrovně návštěvy jednotlivých stránek. Většinu veličin tvoří kategoriální data. Pro řešení úlohy segmentace dataset (viz Obrázek 19: Ukázka upraveného datasetu, který je použitý ve fázi modelování) obsahuje pouze následující veličiny: návštěvník, počet návštěv, počet navštívených stránek během návštěvy, celkový čas na webu, průměrně strávený čas na webu během jedné návštěvy, průměrný počet prohlédnutých stránek během jedné návštěvy a údaj o počtu transakcí. Tento dataset obsahuje 6 veličin s numerickými hodnotami a jednu veličinu jako dimenzi - návštěvník.

Podobný dataset je použit i pro metodu klasifikace uživatelů do skupin s využitím techniky rozhodovacích stromů pracující s numerickými veličinami, avšak s rozdílem, že

místo počtu transakcí je vytvořena veličina s binárními hodnotami 0 nebo 1, která vyjadřuje, zda uživatel provedl transakci, či ne.

Navstevnik	PocetNavstev	PocetNavstivenychStranek	AvgStranek	DobaNavstevy	AvgDobaNavstevy	PocetTransakci
0000010278554503158	6	48	8.0000	1164	194.0000	0
0000020424342248747	9	117	13.0000	2673	297.0000	0
0000027376579751715	4	20	5.0000	196	49.0000	0
0000040862739425590	2	6	3.0000	70	35.0000	0
000005103959234087	4	32	8.0000	808	202.0000	0
0000068403966359845	1	2	2.0000	18	18.0000	0
0000168159078983594	19	1064	56.0000	14649	771.0000	0
0000174067426171406	13	245	18.8462	9906	762.0000	0
0000197671390269035	1	1	1.0000	0	0.0000	0
0000213131142648941	4	52	13.0000	1088	272.0000	1
000026722803385797	3	6	2.0000	54	18.0000	0
0000291342601222013	1	3	3.0000	16	16.0000	0
0000436683523507380	5	34	6.8000	778	155.6000	0
0000458812883559498	1	1	1.0000	0	0.0000	0

Obrázek 19: Ukázka upraveného datasetu, který je použitý ve fázi modelování

Pro fázi interpretace je vytvořen rozšířený dataset obsahující převážně numerické veličiny. Rozšířený dataset obsahuje veličiny návštěvník, jeho počet návštěv, počet navštívených stránek během návštěvy, celkový čas na webu, průměrně strávený čas na webu během jedné návštěvy, průměrný počet prohlédnutých stránek během jedné návštěvy, počet transakcí, koeficient počtu transakcí za měsíc, celkové Revenue, průměrné Revenue, minimální a maximální Revenue, a veličiny obsahující celkové Revenue pro jednotlivé kategorie produktů (jeden sloupec - jedna kategorie produktů), zařízení a zdroj návštěvy. Rozšířený dataset k interpretaci (viz Obrázky 20 a 21: Ukázka rozšířeného datasetu, který je použitý ve fázi interpretace) obsahuje 30 metrik s numerickými hodnotami a jednu veličinu jako dimenzi - návštěvník. Dimenze návštěvník je použita jako klíč pro přiřazení shluků získaných pomocí techniky shlukování k hodnotám rozšířeného datasetu za účelem kompletní interpretace charakteristik jednotlivých shluků.

MaxRevenue	KoefTransakceMesic	Obleceni	Doplnky	Elektro	KancelarskePotreby
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
39.59	0.0833	39.59	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00
0.00	0.0000	0.00	0	0	0.00

KanalReferral	KanalAffiliates	KanalDisplay	KanalOstatni	ZarizeniDesktop	ZarizeniMobil
0.00	0	0	0	0.00	0
0.00	0	0	0	0.00	0
0.00	0	0	0	0.00	0
0.00	0	0	0	0.00	0
0.00	0	0	0	0.00	0
0.00	0	0	0	0.00	0
0.00	0	0	0	0.00	0
0.00	0	0	0	0.00	0
0.00	0	0	0	0.00	0
0.00	0	0	0	39.59	0
0.00	0	0	0	0.00	0
0.00	0	0	0	0.00	0
0.00	0	0	0	0.00	0
0.00	0	0	0	0.00	0

Obrázek 20 a 21: Ukázka rozšířeného datasetu, který je použitý ve fázi interpretace

Důvodem, proč není použit rozšířený dataset i ve fázi modelování je ten, že takto postavený dataset, z důvodu malého počtu transakcí a následně malých výskytů Revenue ve sloupcích s kategoriemi, zařízením a kanály, obsahuje velké množství nulových hodnot, což může mít za následek případné zkreslení vzdáleností při aplikaci techniky shlukování K-středů.

4.3.2. Čištění dat

V rámci fáze porozumění a přípravy dat jsou pro úlohu segmentace upraveny jednotlivé veličiny tak, aby vyjadřovaly stejný obsah pomocí numerických hodnot. Během tohoto kroku vznikly prázdné hodnoty (u absence transakcí), které jsou plošně nahrazeny nulou z důvodu vstupních požadavků pro aplikaci algoritmu ve fázi modelování dat.

4.3.3. Sestavení datasetu

Pro sestavení datasetu pro segmentaci a klasifikaci, kde jeden řádek odpovídá jednomu uživateli, jsou využity agregační funkce počet (například počet návštěv), suma (celková doba strávená na webu) a průměr (průměrný počet prohlédnutých stránek na jednu návštěvu) u uživatelů, kteří opakovaně navštěvovali stránky. Pro sestavení datasetu pro fázi interpretace jsou z iniciačního datasetu transponovány dimenze kategorie produktu, zdroj návštěvy a zařízení, kdy každá z těchto kategorií má vlastní sloupec s konkrétní hodnotou Revenue.

4.3.4. Integrace dat

Přestože primární dataset obsahuje data z více zdrojů (transakční primární data z Google Merchandise Store a data webové analytiky z Google Analytics 360), jsou uložena v Google BigQuery v jedné tabulce. Za účelem vytvoření datasetu pro segmentaci a klasifikaci jsou data agregována dle dimenze návštěvník. Za účelem vytvoření rozšířeného datasetu pro fázi interpretace je iniciační dataset rozdělen do více dočasných tabulek (z důvodu limitu dostupného výpočetního výkonu), nad kterými jsou provedeny transformace a tyto jsou následně spojené do jedné. Během těchto operací jsou data agregována dle dimenze návštěvník a tato dimenze je primárním klíčem pro opětovné spojení dat do jedné tabulky pro fázi interpretace.

4.3.5. Formátování dat

Data pro segmentaci a klasifikaci jsou formátována do jednotných numerických hodnot, není aplikováno žádné specifické řazení pro fázi modelování a ani pro fázi interpretace.

Z podkapitoly vyplývá, že iniciační dataset je transformován pro fázi modelování tak, aby vyjadřoval stejný obsah pomocí numerických hodnot. Během fáze přípravy dat vznikl také rozšířený dataset, který je použit ve fázi interpretace. Při přípravě rozšířeného datasetu vzniklo velké množství prázdných hodnot, které jsou vzhledem k volbě algoritmu nahrazeny nulou. Rozšířený dataset, který obsahuje velké množství nulových hodnot, není pro fázi modelování vhodný z důvodu možného zkreslení vzdáleností při aplikaci techniky shlukování K-středy, avšak v pozdější fázi interpretace, s přiřazenými shluky získanými ve fázi modelování přes dimenzi návštěvník, přináší zajímavé souvislosti.

4.4. Modelování

Podkapitola popisuje fázi modelování, výběr vhodné data miningové techniky pro úlohu segmentace a úlohu klasifikace, design modelů a aplikaci modelovacích technik v platformě RapidMiner s výstupy z obou úloh.

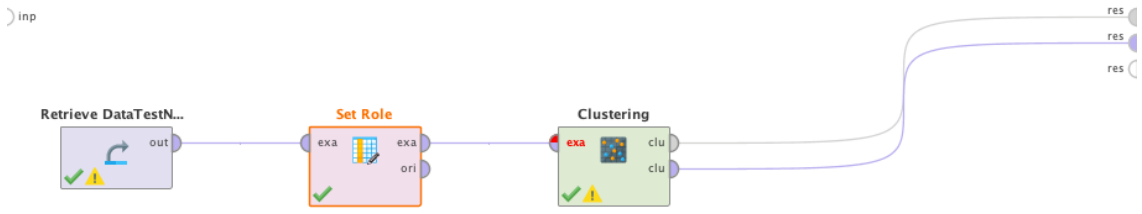
4.4.1. Výběr modelovací techniky

K řešení úlohy segmentace pomocí shlukování je zvolena technika K-středů. U této techniky nejsou povoleny prázdné hodnoty, tyto byly nahrazeny ve fázi přípravy dat nulami.

K řešení úlohy klasifikace je vybrána technika rozhodovací stromy pracující s numerickými veličinami, jejímž cílovým atributem je binární hodnota 0 nebo 1 vyjadřující, zda uživatel provedl transakci, či ne. Pro účely klasifikační úlohy jsou data rozdělena na trénovací a testovací. 80% dat jsou trénovací data a 20% dat jsou data testovací.

4.4.2. Vytvoření testového návrhu

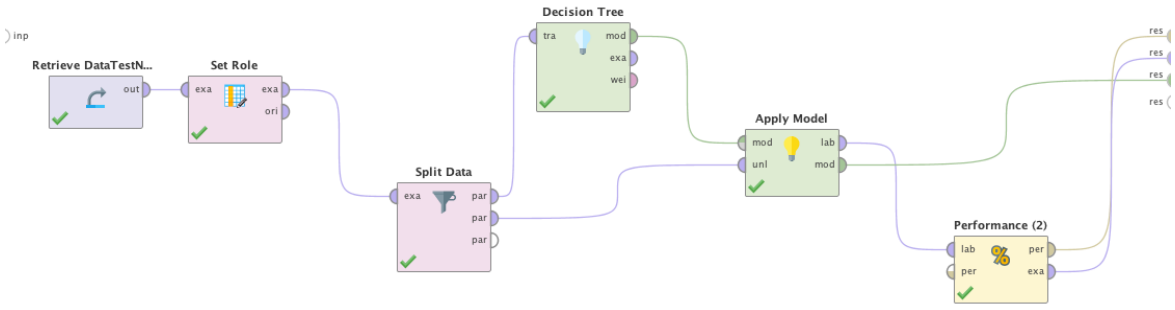
Pro aplikaci techniky K-středů pro řešení úlohy segmentace je využitá platforma RapidMiner. Design modelu (viz Obrázek 22: Design modelu úlohy segmentace s využitím metody K-středů v platformě RapidMiner) obsahuje vstupní data, upravená ve fázi přípravy dat dle požadavků algoritmu, následuje krok nastavení role veličin a v posledním kroku samotná aplikace metody shlukování pomocí techniky K-středů.



Obrázek 22: Design modelu úlohy segmentace s využitím metody K-středů v platformě RapidMiner

Úloha klasifikace je v platformě RapidMiner řešena pomocí techniky rozhodovacích stromů. Design modelu (viz Obrázek 23: Design modelu úlohy klasifikace s využitím metody rozhodovací stromy v platformě RapidMiner) obsahuje vstupní data, která jsou téměř shodná jako při aplikaci metody segmentace. Jediným rozdílem je záměna veličiny počet transakcí za veličinu s binární hodnotou transakce, která určuje, zda byla provedena transakce, či ne. V kroku nastavení rolí je jako cílový atribut nastaven údaj, zda byla provedena transakce či ne. Následuje rozdělení dat na dvě části, kdy trénovací data tvoří 80% dat a testovací data 20% dat. Data jsou algoritmem rozdělena náhodně, zároveň jsou

reprezentativní vůči celku. Poté probíhá aplikace techniky rozhodovacích stromů. Získaný klasifikační model je aplikován na testovací data a je vypočtena přesnost klasifikačního modelu.



Obrázek 23: Design modelu úlohy klasifikace s využitím metody rozhodovací stromy v platformě RapidMiner

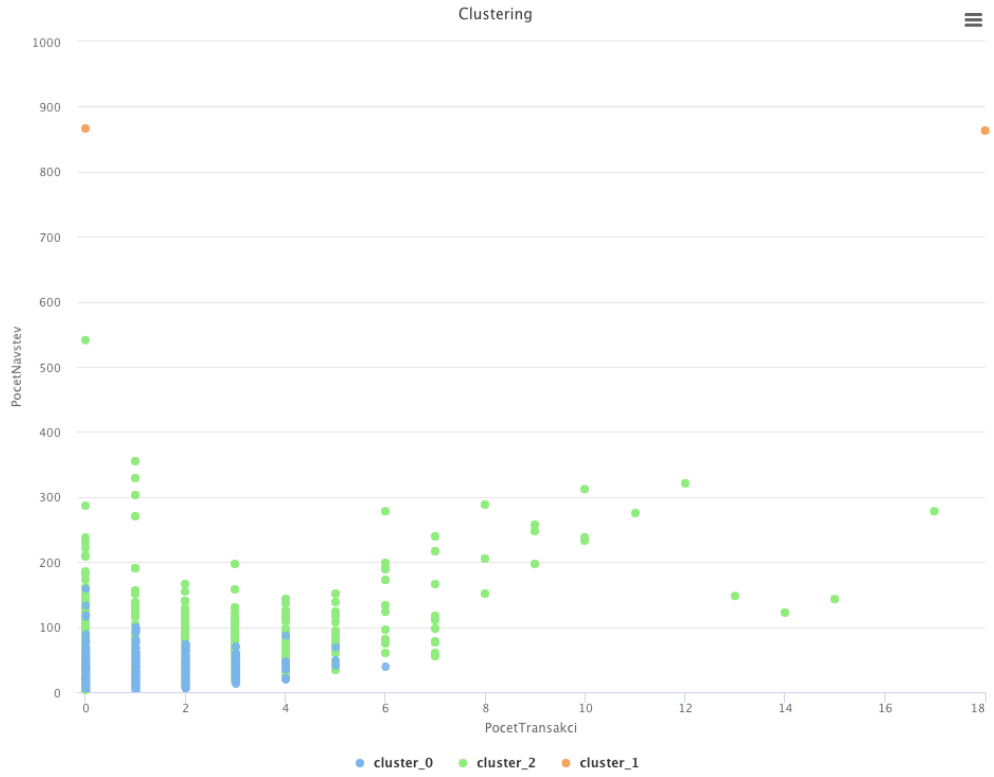
4.4.3. Postavení modelu

Pro postavení modelu pro úlohu segmentace je využit design modelu z předchozího kroku a jsou nastaveny dílčí parametry. V kroku nastavení rolí je veličina návštěvník označena jako dimenze a byla přidána váha veličině počet transakcí. Metoda K-středů vyžaduje určení počtu shluků na vstupu. Počet je určen na tři s maximálním opakováním přepočtu centroidů na 20. V parametrech je dále přidána na výstup dimenze, která označuje shluk, do které je návštěvník zařazen. Vzdálenosti jsou určovány pomocí Euklidovské vzdálenosti.

Pro postavení modelu pro úlohu klasifikace je využit design modelu z předchozího kroku. Jako cílový atribut je nastavena veličina transakce s hodnotami 0 nebo 1, která označuje, zda byla provedena transakce, či ne. Data jsou rozdělena na trénovací a testovací, kdy 80% dat jsou trénovacími daty a 20% testovacími. Kritériem pro dělení dat do podmnožin je Gini index. Maximální hloubka stromu je 5. Je aplikováno prořezávání stromu. Minimální počet příkladů v listové podmnožině je stanoven na 10. Minimální počet příkladů v uzlu je stanoven na 20. Aplikační parametry při aplikaci klasifikačního modelu nastaveny nejsou. Pro zhodnocení úspěšností klasifikačního modelu je zvolen výpočet míry přesnosti při klasifikaci.

4.4.4. Posouzení modelu

Výstup úlohy segmentace, která je řešená technikou K-středů, obsahuje tři shluky. První shluk, pod označením cluster_0, obsahuje většinu návštěvníků, a to 356 145. Druhý shluk, pod označením cluster_1, obsahuje pouze 2 návštěvníky a třetí shluk, pod označením cluster_2, 3357 návštěvníků.

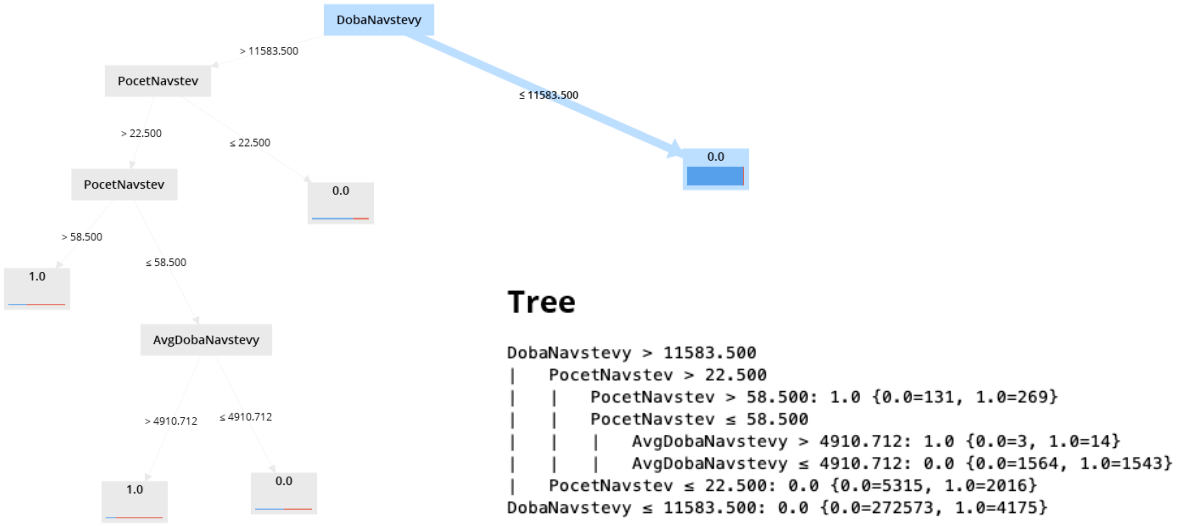


Obrázek 24: Výstup úlohy segmentace s využitím metody K-středů v platformě RapidMiner - na ose x počet transakcí, na ose y počet návštěv

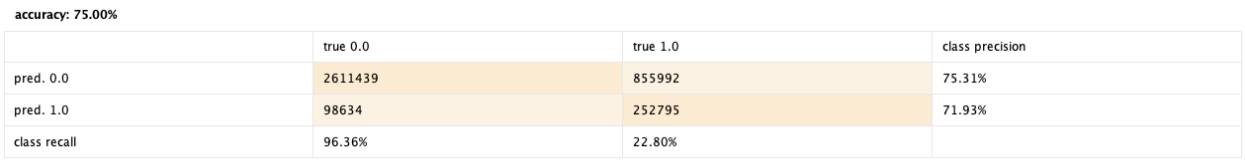
Z prvotního vizuálního posouzení lze vyčíst (viz Obrázek 24: Výstup úlohy segmentace s využitím metody K-středů v platformě RapidMiner - na ose x počet transakcí, na ose y počet návštěv a Obrázek 25: Výstup úlohy segmentace s využitím metody K-středů v platformě RapidMiner - na ose x doba návštěvy, na ose y počet návštěv), že při aplikaci modelu jsou data rozdělena do tří skupin, které mají různé charakteristiky. Z prvotní analýzy dále vyplývá hypotéza, že skupina pod označením cluster_2 může být skupinou uživatelů, kteří reaguji na marketingovou komunikaci častěji, než ostatní.

Obrázek 25: Výstup úlohy segmentace s využitím metody K-středů v platformě RapidMiner - na ose x doba návštěvy, na ose y počet návštěv

Výstupem úlohy klasifikace je rozhodovací strom s pravidly (viz Obrázek 26: Klasifikační model vytvořený v platformě RapidMiner: Rozhodovací strom a pravidla rozhodovacího stromu), ze kterých vyplývá, že klíčovou veličinou, umístěnou v kořeni stromu, je celková doba návštěvy. Již první rozdělení uživatelů do podmnožin významně oddělilo uživatele na dvě skupiny. Většinová skupina obsahuje převážně ty uživatele, kteří neprovedli transakci. Druhá skupina jsou uživatelé, kteří na webu tráví hodně času během opakovaných návštěv a je u nich větší pravděpodobnost uskutečněné transakce. Přesnost tohoto klasifikačního modelu je 75% (viz Obrázek 27: Klasifikační model vytvořený v platformě RapidMiner: Přesnost klasifikačního modelu).



Obrázek 26: Klasifikační model vytvořený v platformě RapidMiner: Rozhodovací strom a pravidla rozhodovacího stromu



accuracy: 75.00%

	true 0.0	true 1.0	class precision
pred. 0.0	2611439	855992	75.31%
pred. 1.0	98634	252795	71.93%
class recall	96.36%	22.80%	

Obrázek 27: Klasifikační model vytvořený v platformě RapidMiner: Přesnost klasifikačního modelu

Z podkapitoly vyplývá, že pro řešení úlohy segmentace je vybrána shlukovací technika K-středy. Pro úlohu klasifikace je vybrána technika rozhodovací stromy pracující s numerickými veličinami, kdy cílovým atributem je veličina s binárními hodnotami 0 nebo 1, která vyjadřuje zda byla provedena transakce, či ne. Design modelu shlukování a klasifikace má na vstupu datasety z fáze přípravy dat. Výstup z úlohy segmentace obsahuje 3 shluky. Je hypotéza, že jeden ze shluků může být skupinou, která reaguje na marketingovou komunikaci častěji, než ostatní. Klasifikační model, který určí, zda uživatel provede transakci, či ne, má přesnost 75%.

Kapitola popisuje fáze a jejich kroky ve shodě s metodologií CRISP-DM až do fáze modelování. V kapitole jsou popsány cíle z obchodního a data miningového hlediska, provedeno porozumění problematice, porozumění datům, jejich příprava a následně aplikace data miningových technik. Následující kapitola na tyto fáze navazuje ve shodě s metodologií CRISP-DM a věnuje se fázím interpretace a využití.

5. Zhodnocení praktické aplikace vybraných metod DM dle metodologie CRISP-DM

Kapitola navazuje na předcházející fáze metodologie CRISP-DM fázemi interpretace a využití. V kapitole je popsáno zhodnocení výsledků, posouzení procesu, další kroky a využití. Závěrečnými kroky fáze využití a zároveň i závěrečnými kroky celého procesu je vypracovaná závěrečná zpráva, která má charakter zprávy pro zadavatele a shrnutí projektu, což je shrnutí doporučení a zkušeností získaných v dané iteraci procesu.

5.1. Interpretace

Podkapitola popisuje zhodnocení výsledků obou data miningových technik, které byly aplikovány ve fázi modelování. Výsledky úlohy segmentace jsou dány do souvislosti s výstupy z popisné statistiky provedené v rámci fáze porozumění datům. Následně je provedeno posouzení procesu ať už z pohledu samotné volby modelovacích technik, tak i z pohledu stanovení cílů na počátku procesu. V posledním kroku této fáze jsou určeny další kroky ať už z pohledu data miningu, tak i v teoretické rovině z obchodního pohledu.

5.1.1. Zhodnocení výsledků

Během zhodnocení výsledků je důležité brát v potaz, že se jedná o vzorová data Google Analytics. Data zahrnují transakce z několika e-shopů z celého světa za období jednoho roku. Z tohoto pohledu je nutné brát výsledky pouze jen jako orientační a slouží spíše jako demonstrace možností praktického využití technik DM v oblasti marketingu. Nelze z nich odvodit obecné zákonitosti.

Ze zjištění popisné statistiky ve fázi porozumění datům vyplývá hypotéza, že skupina, co reaguje na marketingovou komunikaci častěji než ostatní, existuje, a lze u ní rozpoznat určité charakteristiky. Pro tuto skupinu lze vytvořit specifické marketingové

strategie u dalších plánovaných kampaní a tím dosáhnout potencionálního zvýšení konverzního poměru. Ve fázi porozumění datům je určena hypotéza, že skupina, která reaguje častěji na marketingovou komunikaci jsou uživatelé se zájmem o oblečení, kabelky, tašky a kancelářské potřeby, se kterými je komunikováno pomocí přímého marketingu či na web chodí přes marketingové referenční odkazy. Tito uživatelé nakupují kontinuálně v průběhu celého roku, v průměru několikrát za měsíc, mají vysoký konverzní poměr, jejich jednotlivé objednávky nemají vysokou hodnotu a nejčastěji k takovým nákupům využívají desktopové zařízení. Na webu stráví přibližně 20 min, během kterých navštíví kolem 35 stránek.

Za účelem zhodnocení výsledků z fáze modelování, úlohy shlukování, jsou výsledné clustery na základě dimenze návštěvník propojeny s rozšířeným datasetem, obsahující údaje o chování uživatele na webu, tedy s daty měřící jejich zájmy dle kategorií produktů, preferované zařízení či kanál. Je vyhodnoceno, že shluk pod označením cluster_2 obsahuje pouze 3357 návštěvníků, což je 1% z celkového počtu návštěvníků, avšak Revenue za tuto skupinu návštěvníků je přes 5,6 milionů, kdy se jedná o 53,5% z celkového Revenue. Jedná se tedy o velice významnou skupinu uživatelů, v rámci které sice ne všichni zatím nakoupili, ale mají podobné charakteristiky jako ti, co provedli alespoň jednu transakci a nakupují pravidelně. Je tedy pravděpodobné, že transakci udělají i tito uživatelé, kteří zatím transakci neudělali. Tato skupina uživatelů navštíví web v průměru 3 krát za měsíc. Konverzní poměr je u této skupiny 0,02%. Hodnota jedné transakce je v průměru 843, což je výrazně vyšší hodnota oproti hodnotě 9 u uživatelů ve shluku pod označením cluster_0. Na webu stráví v průměru 35 minut a navštíví 43 stránek. Největší zájem mají o kategorie oblečení, kabelky a tašky a kancelářské potřeby. Největší část uživatelů přichází z kanálu Direct a přes marketingové referenční odkazy. Preferují desktopové zařízení. Tato charakteristika skupiny je z určité části podobná charakteristice skupiny zákazníků, která byla stanovená hypotézou na základě popisné statistiky.

Shluk pod označením cluster_0 obsahuje 356145 návštěvníků, což je 99% z celkového počtu. Uživatelé v této skupině průměrně navštěvují web 0,3 krát za měsíc. Konverzní poměr skupiny je 0,007%. Průměrně stráví na webu 3 minuty a navštíví 9 stránek. Přestože nejvíce preferují marketingové referenční odkazy a kanál Direct, ke konverzi dochází v minimu případech a když už ano, tak mají spíše nižší hodnoty. Oblíbenými kategoriemi zboží je oblečení, kancelářské potřeby a kabelky a tašky, což je v korelaci s

výstupy popisné statistiky. Preferují desktopové zařízení. Je zajímavým jevem, že přestože se jedná o výraznou většinu uživatelů, jejich podobné charakteristiky zapříčinily to, že ve vizuálním vyjádření (viz Obrázek 25: Výstup úlohy segmentace s využitím metody K-středů v platformě RapidMiner - na ose x doba návštěvy, na ose y počet návštěv), tvoří oproti clusteru_2 vizuální menšinu.

Shluk pod označením cluster_1 obsahuje pouze 2 návštěvníky, kdy jednoho z nich lze charakterizovat jako významného velkoodběratele zboží v kategoriích oblečení, kabelky a tašky, sklenice a hrnky, elektro a kancelářské potřeby. Nakupovat chodí hlavně z kanálu Direct na desktopu. Celkové Revenue za tohoto návštěvníka je necelých 1,4 milionů za rok, což je 13,2% z celkového Revenue. Průměrně navštěvuje web 70 krát za měsíc, za jednu návštěvu v průměru navštíví 30 stránek a stráví na webu 40 minut. Jeho konverzní poměr je 0,02%.

Výstupem úlohy klasifikace je rozhodovací strom s pravidly (viz Obrázek 26: Klasifikační model vytvořený v platformě RapidMiner: Rozhodovací strom a pravidla rozhodovacího stromu), ze kterých vyplývá, že klíčovou veličinou, umístěnou v kořeni stromu, je celková doba návštěvy. Již prvním rozdělení uživatelů do podmnožin se výrazně odděluje většina uživatelů, u kterých je předpoklad, že transakci neprovedou, od menšiny uživatelů, co ano. Tito uživatelé na webu tráví hodně času během opakovaných návštěv. Přesnost tohoto klasifikačního modelu je 75% (viz Obrázek 27: Klasifikační model vytvořený v platformě RapidMiner: Přesnost klasifikačního modelu). Je zajímavým jevem, že tyto výstupy jsou v korelaci se zjištěními získanými metodou shlukování.

Lze konstatovat, že úloha shlukování nalezla skryté vzory v datech, které rozřadily uživatele do skupin relevantních pro další použití. Stanovené cíle z pohledu data miningu jsou splněny jak pomocí výstupů popisné statistiky nalezením skupiny uživatelů, kteří reagují častěji, než ostatní skupiny na marketingovou komunikaci ve fázi porozumění datům, ale i nalezením skrytých struktur v datech ve fázi modelování, které uživatele uspořádaly do shluků. Tímto je splněn hlavní cíl praktické aplikace i z obchodního hlediska, kdy je nalezena skupina uživatelů, která reaguje na marketingové kampaně častěji, než ostatní. Na tuto skupinu lze zacílit kampaně a následně zhodnotit její úspěšnost z obchodního hlediska pomocí konverzního poměru. Stanovený vedlejší cíl, určení u nového uživatele, zda provede transakci, či ne, lze splnit aplikací klasifikačního modelu s přesností 75%. Míra přesnosti

klasifikačního modelu splňuje na počátku stanovený cíl z pohledu data miningu, který byl určen na 70%.

5.1.2. Posouzení procesu

Proces umožňuje splnění na počátku stanovených obchodních a data miningových cílů. Volba data miningové techniky K-středů pro úlohu segmentace, která je méně náročná na výpočetní výkon, než hierarchické shlukování, umožnila provedení úlohy nad velkým rozsahem dat. Výstup úlohy s využitím zmíněné techniky je relevantní jak z hlediska data miningového, tak i obchodního, a lze ji tak hodnotit jako vhodnou. Data miningová technika rozhodovací stromy pro vytvoření klasifikačního modelu má přesnost 75%. Volba cílového atributu mající binární hodnotu 0 nebo 1 z hlediska výpočetního výkonu umožnila relativně rychlé vytvoření klasifikačního modelu nad velkým rozsahem dat a následnou aplikaci modelu na testovací data. Je věnováno velké množství času fázím porozumění datům a přípravy dat, které jsou provedeny velice důkladně. Tato skutečnost je pro úspěšnost v následující fázi modelování klíčová.

5.1.3. Určení následujících kroků

Z obchodního hlediska jsou dalšími kroky tvorba obchodních a marketingových strategií a následné nasazení kampaní pro nalezené skupiny uživatelů. Vyhodnocením jejich úspěšnosti pomocí konverzního poměru lze sledovat míru splnění cílů z obchodního hlediska, která je zatím splněna teoreticky. Je možné porovnat úspěšnost zacílení na skupiny rozřazené pomocí popisné statistiky ve fázi porozumění datům a na skupinu nalezenou ve fázi modelování pomocí techniky shlukování.

Z pohledu data miningu je dalším krokem nová iterace celého procesu poté, co proběhnou výše zmíněné další kroky z obchodního hlediska. Cílem je získat přesnější klasifikační model a znovu provést úlohu segmentace na podkladě výsledků z nasazení kampaní při zacílení na skupiny nalezené v aktuální iteraci.

Ze zhodnocení výsledků vyplývá, že na počátku stanovené cíle jak z obchodního, tak z data miningového hlediska, jsou v rámci procesu splněny. Vzhledem k tomu, že se jedná o vzorová data Google Analytics, nelze z těchto výstupů odvodit obecné zákonitosti. Data miningová úloha shlukování nalezla skryté vzory v datech, které rozřadily uživatele do

skupin relevantních pro další použití, kdy jednu z nich lze označit jako skupinu uživatelů, která reaguje na marketingové kampaně častěji, než ostatní. Na základě těchto výsledků lze optimalizovat obchodní a marketingové strategie. Vytvořený klasifikační model s využitím data miningové techniky rozhodovací stromy má přesnost určení u nového uživatele, zda provede transakci, či ne, 75%. Důraz na porozumění problematice, datům a jejich přípravě, v kombinaci s vhodnou volbou data miningových technik ve fázi modelování, umožnilo splnění na počátku stanovené cíle. Dalšími kroky může být nasazení kampaní zacílené na nalezené skupiny uživatelů a na základě vyhodnocené úspěšnosti spuštění nové iterace procesu.

5.2. Využití

Podkapitola obsahuje závěrečnou zprávu, která má charakter shrnutí výsledků procesu ve formě prezentovatelné zadavateli, a shrnutí projektu, jehož účelem je shrnout zkušenosti získané na projektu, posoudit průběh procesu a vytvořit doporučení pro podobné typy úloh řešené v budoucích projektech. V rámci diplomové práce se nepočítá s nasazením do provozu ani s reálnou demonstrací využití zadavateli na praktických ukázkách. Z tohoto důvodu plán nasazení ani plán monitorování a správy není součástí praktické aplikace diplomové práce.

5.2.1. Plán nasazení

Vzhledem k tomu, že v rámci diplomové práci není prováděno žádné praktické ověření výstupů procesu, plán nasazení není součástí práce.

5.2.2. Plán monitorování a správy

Vzhledem k tomu, že v rámci diplomové práci není prováděno žádné praktické ověření výstupů procesu, plán monitorování a správy není součástí práce.

5.2.3. Závěrečná zpráva

Praktická aplikace má z obchodního hlediska za cíl nalézt, pokud existuje, skupinu uživatelů, kteří reagují na marketingové kampaně častěji, než ostatní, a nalézt případné souvislosti v databázi, za účelem optimalizace marketingové a obchodní strategie. Dalším

cílem je schopnost odhadnout, kdo z nově akvirovaných uživatelů provede transakci a tomu od počátku přizpůsobit obchodní a marketingovou komunikaci.

Na základě fáze porozumění problematiky a porozumění datům jsou vytvořeny prvotní hypotézy, které jsou následně ověřeny ve fázi modelování s využitím data miningových technik. Ke splnění cíle nalézt skupinu uživatelů, která reaguje častěji, než ostatní, je využita data miningová technika K-středy, která nalezla skryté vzory v datech, které rozřadily uživatele do skupin relevantních pro další použití. Jedna z nalezených skupin splňuje charakteristiku skupiny, která reaguje na marketingovou komunikaci častěji, než ostatní. Tato skupina, byť tvoří 1% z celkového množství uživatelů, přinesla 53,5% z celkového Revenue. Za účelem vytvoření klasifikačního modelu, který odhadne, kdo z nových uživatelů provede transakci, či ne, je využita technika rozhodovací stromy. Přesnost tohoto klasifikačního modelu je 75%. Již prvním rozdělení uživatelů do dvou podmnožin se na základě délky návštěvy výrazně separuje většina uživatelů, u kterých je předpoklad, že transakci neprovedou. Tento jev je v korelaci s výstupy z úlohy shlukování.

Z výstupů získaných ve fázích modelování a interpretace lze optimalizovat marketingovou a obchodní strategii. Následujícím postupem může být zacílení kampaně na skupiny nalezené jak ve fázi porozumění datům s využití popisné statistiky, tak i na skupinu uživatelů, nalezenou s využitím data miningových technik a porovnat úspěšnost. Indikátorem pro vyhodnocení úspěšnosti z obchodního hlediska může být konverzní poměr. Tento výstup může být vstupem pro následující iteraci a optimalizaci *KDD* procesu.

Součástí závěrečné zprávy v rámci metodologie bývá závěrečná prezentace zadavateli, tato však není součástí praktické aplikace diplomové práce.

5.2.4. Shrnutí projektu

Řešení projektu a dílčí rozhodnutí během procesu jsou přímo ovlivněny mojí dosavadní pracovní zkušeností a výstupy z případových studií, které jsou součástí diplomové práce. Díky výše uvedenému se volba data miningových technik, které jsou aplikovány ve fázi modelování, ukázala jako vhodná. Data miningové techniky pro úlohu shlukování, technika K-středy, vzhledem k charakteru techniky, vyžaduje dataset obsahující numerické veličiny. Je nutné volit data tak, aby dataset neobsahoval velké množství nulových hodnot,

jelikož tato skutečnost může mít za následek případné zkreslení vzdáleností při aplikaci metody shlukování a takový výstup pak sice obsahuje shluky, ale při vizuálním vykreslení je zřejmé, že se překrývají a z obchodního hlediska nejsou relevantní. Za těchto okolností pak nelze na základě charakteristik uživatelů separovat skupinu, která reaguje na marketingovou komunikaci častěji, než ostatní. Během procesu se v této souvislosti projevil iterativní charakter metodologie *CRISP-DM*, především mezi fázemi příprava dat a modelování. Důležitou zkušeností je, že byť byly v prvních iteracích z formálního hlediska požadavky modelovací techniky splněny, výstupy byly z obchodního hlediska nerelevantní. Relevantních výsledků se dosáhlo až po opakovaných iteracích. Opakované iterace mezi fázemi příprava dat a modelování proběhly i během úlohy klasifikace, kdy až volba aktuálního cílového atributu v binárním formátu měla za následek relevantní výstup, tedy vytvoření klasifikačního modelu s přesností 75%.

Zajímavou zkušeností je zjištění, že byť mohou data u některých uživatelů působit na první pohled jako anomálie, je možné u nich nalézt zajímavé souvislosti, například skutečnost, že se jedná o velkoodběratele konkrétních kategorií chodící na web z kanálu přímého marketingu.

V podkapitole jsou popsány výsledky praktické aplikace jako výstup zadavateli, které umožňují optimalizovat marketingovou a obchodní strategii. Jedná se o praktickou ukázku využití data miningových technik v oblasti marketingu. Vzhledem ke zdroji dat se jedná pouze o příklad a výstupy nemají charakter obecných zákonitostí. Podkapitola dále popisuje důležité zkušeností získané během praktické aplikace, obzvlášť skutečnost, že přestože je již na počátku zvolena vhodná data miningová technika, relevantních výstupů z obchodního hlediska může být dosaženo až po opakovaných iteracích mezi fázemi příprava dat a modelování. Výběr vhodných data miningových technik na základě vlastních praktických zkušeností a případových studií uvedených v diplomové práci, které mají podobné obchodní cíle, jsou pro dosažení cílů klíčové.

Fáze interpretace a využití jsou závěrem metodologie *CRISP-DM* a praktické aplikace. Splnění stanovených cílů praktické aplikace jak z pohledu data miningu, tak z obchodního hlediska, umožňuje optimalizaci marketingové a obchodní strategie. Klasifikační model s přesností 75% umožňuje rozpoznat u nového uživatele, zda provede transakci, či ne, a přizpůsobit tomu další komunikaci. S 1% uživateli, kteří pokrývají 53,5%

Revenue, lze pomocí následné personalizace podpořit a udržovat dlouhodobé vztahy v rámci speciální obchodní strategie. Proces přináší zajímavá zjištění a zkušenosti, které lze u budoucích projektů, které mají podobné obchodní cíle, využít.

6. Závěr

Data mining nelze nikdy vydělit jako nezávislou disciplínu. Vždy navazuje na fázi před, a to předzpracování dat, a na fázi po, jejich interpretaci. Tento proces tvoří celek *Dobývání znalostí z databází*, kdy při vynechání jedné z fází by nebylo možné validní a relevantní znalost získat. V marketingu, ale i v jiných oblastech, *KDD* proces nemá jasně určený konec. Tuto skutečnost předurčuje již sám o sobě životní cyklus procesu, kdy výstup z jednoho cyklu je zároveň vstupem pro další. Je zdrojem pro další tvorbu hypotéz, zpřesňování otázek a hledání dalších odpovědí. Důvod nesplněných očekávání nemusí být vždy jen ten, že bylo navrženo špatné řešení. Znovu se dostáváme na začátek životního cyklu procesu - zda byly dílčí otázky položeny správně. Je to zároveň příležitost položit dílčí otázky lépe, či upřesnit oblast dat pro dotazování.

V oblasti marketingu bývá ve většině případů z obchodního hlediska cílem personalizace reklamy a udržení si stávajících zákazníků. Důvodem je aktuální trend zákaznické zkušenosti, který se v posledních letech vyvíjel v důsledku rostoucího vlivu sociálních sítí a stále větší dostupnosti informací. Udržení si loajálního zákazníka přináší společnosti stabilní zisk, který je v důsledku mnohem větší, než zisk z nově akvirovaného zákazníka. Ke splnění těchto obchodního cílů lze využít data miningové techniky, kdy pro úlohy segmentace například techniky shlukování a pro úlohy klasifikace techniky rozhodovací stromy. Tyto techniky jsou v oblasti marketingu jedny z často využívaných.

Při aplikaci data miningových technik v rámci *KDD* procesu lze využít nezávislou a univerzálně použitelnou metodologii *CRISP-DM*, která reflektuje zkušenosti a praxi z reálného světa a detailně popisuje logiku celého *KDD* procesu, kterou lze aplikovat na jakýkoli projekt a technologii. V diplomové práci jsou popsány vybrané data miningové techniky, kdy některé z nich jsou následně i součástí fáze modelování, vhodné k použití v oblasti marketingu. Jedná se o techniky shlukování, rozhodovací stromy, asociační pravidla či regresní analýza. Popsány jsou také techniky deskripce a agregace dat, které jsou vhodnými technikami v rámci fáze porozumění datům.

V případových studiích jsou řešeny podobné cíle jako byly stanoveny v praktické aplikaci. Ve všech případech se jedná o aplikaci data miningových technik v marketingové oblasti s cílem personalizace reklamy a udržení dlouhodobých vztahů se zákazníků. Tyto cíle jsou podobné stanoveným cílům v praktické aplikaci. Výstupy z případových studií v kombinaci s vlastními praktickými zkušenostmi ovlivnily výběr technik pro úlohy segmentace a klasifikace řešené v rámci praktické části diplomové práce.

V praktické aplikaci jsou stanovené cíle z obchodního a data miningového hlediska splněny. Je nalezena skupina uživatelů, která reaguje na marketingovou komunikaci častěji, než ostatní. Tato skupina je nalezena jak ve fázi porozumění datům pomocí popisné statistiky, tak i pomocí data miningové techniky shlukování, kdy jsou nalezeny skryté vzory v datech ve fázi modelování. Na tuto skupinu lze zacílit kampaně a následně zhodnotit její úspěšnost z obchodního hlediska pomocí konverzního poměru. S využitím techniky rozhodovacích stromů je vytvořen klasifikační model, který s přesností 75% určí, zda nový návštěvník provede transakci, či ne.

V závěru práce si dovolím konstatovat, že data miningové techniky jsou v oblasti marketingu vhodnými nástroji k získání konkurenční výhody. Neustálý vývoj v oblasti *Dobývání znalostí z databází*, trvalý růst výpočetního výkonu a jeho zvyšující se dostupnost v kombinaci s vývojem v oblasti marketingu vytváří velký potenciál pro vytvoření velmi mocného prostředku k dosažení obchodních cílů společností.

7. Seznam použité literatury

ABBAS, Safia, 2015. Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset. *International Journal of Computer Applications* [online]. Vol. 110 Iss 3, 1-7p. [cit. 2019-08-13] DOI: 10.5120/19293-0725. ISSN 09758887. Dostupné z: http://research.ijcaonline.org/volume110/number3/pxc3900725.pdf

ALEXANDRE, C., et al., 2018. *Marketing Behaviors Analysis in a Mobile Wallet Solution Using Data Mining.* Proceedings [online] - 9th International Conference on Computational Intelligence and Communication Networks, CICN 2017. 2018-January, 88 - 92p. [cit. 2019-10-02] DOI: 10.1109/CICN.2017.8319362. ISBN 9781509050017.

BELLINGER, Gene; CASTRO, Durval; MILLS, Anthony, 2004. *Data, information, knowledge and wisdom* [online] [cit. 2019-08-22]. Dostupné z: http://www.systems-thinking.org/dikw/dikw.htm

BERKA, Petr, 2003. *Dobývání znalostí z databází*. Praha: Academia. Vyd. 1. 366 s. ISBN 80-200-1062-9.

BERNSTEIN, Jay H. The Data-Information-Knowledge-Wisdom Hierarchy and its Antithesis. *NASKO* [online]. 2009, Vol. 2, issue 1, p. 68-75 [cit. 2019-09-15]. DOI: 10.7152/nasko.v2i1.12806. ISSN 2311-4487. Dostupné z: http://journals.lib.washington.edu/index.php/nasko/article/view/12806

CHAPMAN, Pete, et al., 2000. CRISP-DM 1.0: Step-by-step data mining guide. *The Modeling Agency* [online]. Pittsburgh: One Oxford Centre, Copyright © SPSS Inc. [cit. 2019-06-17] Dostupné z: https://www.the-modeling-agency.com/crisp-dm.pdf

CHIK, Paul, 1997. What is Data mining and how to apply data mining techniques to exploit information from your data warehouse or data mart. *Data Mining, Data Warehousing & Client/Server Databases: proceedings of the 8th International Database Workshop* (industrial volume)*, Hong Kong, 29-31 July 1997*. Singapore: Springer, 94 - 102 s. ISBN 981-3083-53-0.

CORIC, D. et al., 2015. Customer Segmentation Using Data Warehouse and Neural Networks. *WSEAS Transactions on Business and Economics* [online]. 186-197p. [cit. 2019-09-15] E-ISSN: 2224-2899. Dostupné z:http://www.wseas.org/multimedia/journals/economics/2015/a345807-093.pdf

FAYYAD,U., et al., 1996. Knowledge discovery and data mining: Towards a unifying framework. *KDD-96 Proceedings: Second International Conference on Knowledge Discovery & Data Mining* [online]. Menlo Park, CA: AAAI Press. 82–88 s. [cit. 2019-07-23] Dostupné z: https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf

GREENWALD, R. 2018. *Beyond the Data Warehouse: New Data Management for Analytics*. GARTNER, Webinar [online] [cit. 2019-07-23]. Dostupné z: https://www.gartner.com/en/webinars/3887381/beyond-the-data-warehouse-new-data-management-for-analytics

GROMOV, V., 2012. Data mining techniques in Real-time Marketing. *Proceedings of the Spring/Summer Young Researchers 19 Colloquium on Software Engineering* [online]. ISSN 23117230. [cit. 2019-09-15] Dostupné z: http://syrcose.ispras.ru/2012/files/submissions/33_syrcose2012_submission_33.pdf

JANČA, J., 2018. *Proč selhávají korporátní BigData a AI projekty?* Umělá inteligence: Komu pomůže a koho zničí? Přednáška na konferenci New Media Inspiration. Praha, ČVUT.

LINOFF, Gordon S., BERRY, Michael J.A., 2011. *Data mining techniques: for marketing, sales, and customer relationship management*. Wiley Pub., Indianapolis, 3rd ed. 847 p. ISBN 978-111-8087-503.

MICROSOFT, 2019. *Co je digitální transformace?* Microsoft digital transformation [online] [cit. 2019-09-15]. Dostupné z: https://www.microsoft.com/cs-cz/digitaltransformation/default.aspx

MICROSOFT, 2019. *Identify job roles.* Azure for the Data Engineer, Microsoft docs e-learn modules [online] [cit. 2019-09-15]. Dostupné z: https://docs.microsoft.com/en-us/learn/modules/data-engineering-processes/2-roles-and-responsibilities

OREŠKI, D. a N. BEGIČEVIĆ REĐEP, 2018. Data-driven decision-making in classification algorithm selection. *Journal of Decision Systems* [online]. Vol. 27:sup 1, 248 - 255 s. [cit. 2019-09-15] DOI: 10.1080/12460125.2018.1468168. ISSN 21167052. Dostupné z:https://doi.org/10.1080/12460125.2018.1468168

PAVLOVIC, D. et al, 2014. Application of data mining in direct marketing in banking sector. *Industrija* [online]. Vol 42, br. 1, 189-201. [cit. 2019-09-15] DOI: 10.5937/industrija42-5087. ISSN 0350-0373. Dostupné z: http://scindeks.ceon.rs/Article.aspx?artid=0350-03731401189P

PEMBERTON, Chris, 2019. *Gartner Predicts 2019: In Search of Balance in Marketing.* GARTNER [online] [cit. 2019-10-15]. Dostupné z: https://www.gartner.com/smarterwithgartner/gartner- predicts-2019-in-search-of-balance-in-marketing/

PIATETSKY-SHAPIRO,G., 2014. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. *KDnuggets.com* [online] [cit. 2019-09-15]. Dostupné z: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

PRŮCHA L., 2010. *Popisná statistika.* Materiály pro studenty [online] [cit. 2019-11-11]. Dostupné z: https://math.feld.cvut.cz/prucha/mstp/7pu.pdf

RALUCA-Cristina, M., 2012. *New approaches to customer base segmentation for small and medium-sized enterprises*. Annals of Faculty of Economics [online], Vol 1, issue 2, p. 848-854. [cit. 2019-11-11] Dostupné z:http://anale.steconomiceuoradea.ro/volume/2012/n2/129.pdf

RUD, Olivia Parr, 2001. *Data mining: praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM)*. Vyd. 1. Praha: Computer Press. 329 s. Rychle a jistě. Databáze. ISBN 80-7226-577-6.

SAS, 2017. *Introduction to SEMMA.* SAS Enterprise Miner 14.3: Reference Help [online]. [cit. 2019-11-11] Dostupné z: https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en

SING'OEI, Lilian a Jiayang WANG, 2013. *Data Mining Framework for Direct Marketing: A Case Study of Bank Marketing*. IJCSI International Journal of Computer Science Issues. Vol. 10 Issue. 2, 198-203p. ISSN 1694-0814.

SKALSKÁ, Hana, 2010. *Data mining a klasifikační modely*. Vyd. 1. Hradec Králové: Gaudeamus, 154 s. Recenzované monografie; 4. ISBN 978-80-7435-088-7.

SKLENÁK, Vilém a kol., 2001. *Data, informace, znalosti a Internet*. C.H. Beck pro praxi. Vyd. 1. Praha: C.H. Beck, 507 s. ISBN 80-717-9409-0.

TAI, Samson, 1997. Data Mining: the software that finds pattern never seen before. *Data mining, data warehousing & client/server databases: Proceedings of the 8th international database workshop (industrial volume), Hong Kong, 29-31 July 1997*. Singapore: Springer, 1997. xii, 303 s. ISBN 981-3083-53-0.

TAN, Pang-Ning, Michael STEINBACH a Vipin KUMAR, 2013. *Introduction to data mining.* Pearson new international edition. Harlow: Pearson Education, 732 s. ISBN 978-1-292-02615-2.

VOICU Mirela-Cristna et al., 2011. *Data Mining – Innovative Method for Obtaining Information in Marketing and Business Management. EIRP Proceedings* [online]. Vol 6, Issue 1, Pp 621-626. [cit. 2019-07-23] ISSN 20679211.

8. Seznam obrázků

9. Seznam zkratek

KDD	Knowledge Discovery in Databases
DM	Data mining
CRISP-DM	Cross-industry standard process for data mining
SEMMA	Akronym pro Sample, Explore, Modify, Model, Assess
5A	Akronym pro Asses, Acces, Analyze, Akt, Automate
CX	Customer Experience
KPI	Key Performance Indicator
CTR	Click through rate
CPC	Cost per click
CPA	Cost per action
CR	Conversion Rate
CRM	Customer Relationship Management
MOOC	Massive Open Online Courses
ZB	Zettabyte

Příloha 1: SQL dotazy použité v praktické aplikaci

SQL dotaz do Google BigQuery přes API s využitím programovacího jazyka R

(zdroj dat: http://bigquery.cloud.google.com

zdroj pro dotaz: https://cloud.google.com/ai-platform/notebooks/docs/use-r-bigquery)

```
library(tidyverse)
library(httpuv)
library(gargle)
library(bigrquery)

projectid = "project-diploma-datamining"

sql <- "SELECT
IFNULL(fullVisitorId,'0') as Navstevnik,
IFNULL(visitId,0) as Navsteva,
date as Datum,
IFNULL(channelGrouping,'0') as ZdrojNavstevy,
IFNULL(device.deviceCategory,'0') AS Zarizeni,
IFNULL(hits.contentGroup.previousContentGroup2,'0') AS KategorieProduktu,
IFNULL(hits.eCommerceAction.action_type,'0') as TypAkce,
IFNULL(totals.timeOnSite,0) AS DobaNavstevy,
IFNULL(totals.pageviews,0) AS PocetNavstivenychStranek,
IFNULL(hits.transaction.transactionId,'0') as TransactionID,
IFNULL(SUM(hits.transaction.transactionRevenue/1000000),0) as Revenue
FROM
`bigquery-public-data.google_analytics_sample.ga_sessions_*`,
UNNEST(hits) AS hits,
UNNEST(hits.product) AS hitsproduct
WHERE
_table_suffix BETWEEN '20160801'
AND FORMAT_DATE('%Y%m%d',DATE_SUB(CURRENT_DATE(), INTERVAL 1 DAY))
GROUP BY 1,2,3,4,5,6,7,8,9,10
ORDER BY 1,2,3,4,5,6,7,8,9,10 DESC
"

df <- query_exec(sql, projectid, use_legacy_sql = FALSE, max_pages=Inf)
```

SQL dotazy použité ve fázi porozumění a přípravy dat

```
-- uživatelé co provedli transakci,transakcni data
CREATE TABLE TransakceAGR as SELECT
    Navstevnik as Navstevnik,
```

```
    COUNT(TransactionID) AS PocetTransakci,
    SUM(Revenue) AS Revenue,
    AVG(Revenue) AS AvgRevenue,
    MIN(Revenue) AS MinRevenue,
    MAX(Revenue) AS MaxRevenue,
    COUNT(TransactionID) / 12 AS KoefTransakceMesic
FROM GA_data
WHERE
    TransactionID <> '0'
GROUP BY 1;

-- uzivatele co provedli transakci - Revenue po produktech
CREATE TABLE ProduktyRevenue as SELECT
    A.Navstevnik as Navstevnik,
    B.Obleceni AS Obleceni,
    C.Doplnky AS Doplnky,
    D.Elektro AS Elektro,
    E.KancelarskePotreby AS KancelarskePotreby,
    F.KabelkyTasky AS KabelkyTasky,
    G.Brand AS Brand,
    H.SkleniceHrnky AS SkleniceHrnky,
    I.ZivotniStyl AS ZivotniStyl
FROM
    GA_data A
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik,
            KategorieProduktu AS KategorieProduktu,
            SUM(Revenue) AS Obleceni
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND KategorieProduktu = 'Apparel'
    GROUP BY 1) B ON A.Navstevnik = B.Navstevnik
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS Doplnky
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND KategorieProduktu = 'Accessories'
    GROUP BY 1) C ON A.Navstevnik = C.Navstevnik
        LEFT JOIN
    (SELECT
```

```
    Navstevnik AS Navstevnik, SUM(Revenue) AS Elektro
FROM
    GA_data
WHERE
    TransactionID <> '0'
        AND KategorieProduktu = 'Electronics'
GROUP BY 1) D ON A.Navstevnik = D.Navstevnik
    LEFT JOIN
(SELECT
    Navstevnik AS Navstevnik, SUM(Revenue) AS KancelarskePotreby
FROM
    GA_data
WHERE
    TransactionID <> '0'
        AND KategorieProduktu = 'Office'
GROUP BY 1) E ON A.Navstevnik = E.Navstevnik
    LEFT JOIN
(SELECT
    Navstevnik AS Navstevnik, SUM(Revenue) AS KabelkyTasky
FROM
    GA_data
WHERE
    TransactionID <> '0'
        AND KategorieProduktu = 'Bags'
GROUP BY 1) F ON A.Navstevnik = F.Navstevnik
    LEFT JOIN
(SELECT
    Navstevnik AS Navstevnik, SUM(Revenue) AS Brand
FROM
    GA_data
WHERE
    TransactionID <> '0'
        AND KategorieProduktu = 'Brands'
GROUP BY 1) G ON A.Navstevnik = G.Navstevnik
    LEFT JOIN
(SELECT
    Navstevnik AS Navstevnik, SUM(Revenue) AS SkleniceHrnky
FROM
    GA_data
WHERE
    TransactionID <> '0'
        AND KategorieProduktu = 'Drinkware'
GROUP BY 1) H ON A.Navstevnik = H.Navstevnik
    LEFT JOIN
(SELECT
    Navstevnik AS Navstevnik, SUM(Revenue) AS ZivotniStyl
```

```
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND KategorieProduktu = 'Lifestyle'
    GROUP BY 1) I ON A.Navstevnik = I.Navstevnik
WHERE
    TransactionID <> '0'
GROUP BY 1;

-- uzivatele co provedli transakci, Revenue po zarizenich
CREATE TABLE ZarizeniRevenue as SELECT
    A.Navstevnik as Navstevnik,
    J.ZarizeniDesktop AS ZarizeniDesktop,
    K.ZarizeniMobil AS ZarizeniMobil,
    L.ZarizeniTablet AS ZarizeniTablet
FROM
    GA_data A
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS ZarizeniDesktop
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND Zarizeni = 'desktop'
    GROUP BY 1) J ON A.Navstevnik = J.Navstevnik
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS ZarizeniMobil
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND Zarizeni = 'mobile'
    GROUP BY 1) K ON A.Navstevnik = K.Navstevnik
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS ZarizeniTablet
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND Zarizeni = 'tablet'
    GROUP BY 1) L ON A.Navstevnik = L.Navstevnik
WHERE
```

```
    TransactionID <> '0'
GROUP BY 1;


-- uzivatele co provedli transakci, Revenue po kanalech
CREATE TABLE KanalyRevenue as SELECT
    A.Navstevnik as Navstevnik,
    M.KanalDirect AS KanalDirect,
    N.KanalOrganicSearch AS KanalOrganicSearch,
    O.KanalPaidSearch AS KanalPaidSearch,
    P.KanalSocial AS KanalSocial,
    Q.KanalReferral AS KanalReferral,
    R.KanalAffiliates AS KanalAffiliates,
    S.KanalDisplay AS KanalDisplay,
    T.KanalOstatni AS KanalOstatni
FROM
    GA_data A
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS KanalDirect
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND ZdrojNavstevy = 'Direct'
    GROUP BY 1) M ON A.Navstevnik = M.Navstevnik
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS KanalOrganicSearch
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND ZdrojNavstevy = 'Organic Search'
    GROUP BY 1) N ON A.Navstevnik = N.Navstevnik
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS KanalPaidSearch
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND ZdrojNavstevy = 'Paid Search'
    GROUP BY 1) O ON A.Navstevnik = O.Navstevnik
        LEFT JOIN
```

```
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS KanalSocial
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND ZdrojNavstevy = 'Social'
    GROUP BY 1) P ON A.Navstevnik = P.Navstevnik
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS KanalReferral
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND ZdrojNavstevy = 'Referral'
    GROUP BY 1) Q ON A.Navstevnik = Q.Navstevnik
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS KanalAffiliates
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND ZdrojNavstevy = 'Affiliates'
    GROUP BY 1) R ON A.Navstevnik = R.Navstevnik
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS KanalDisplay
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND ZdrojNavstevy = 'Display'
    GROUP BY 1) S ON A.Navstevnik = S.Navstevnik
        LEFT JOIN
    (SELECT
        Navstevnik AS Navstevnik, SUM(Revenue) AS KanalOstatni
    FROM
        GA_data
    WHERE
        TransactionID <> '0'
            AND ZdrojNavstevy = '(Other)'
    GROUP BY 1) T ON A.Navstevnik = T.Navstevnik
WHERE
    TransactionID <> '0'
```

```
GROUP BY 1;

-- všichni uživatelé
CREATE TABLE Uzivatele AS SELECT
    Navstevnik,
    COUNT(Navsteva) AS PocetNavstev,
    SUM(PocetNavstivenychStranek) AS PocetNavstivenychStranek,
    SUM(PocetNavstivenychStranek) / COUNT(Navsteva) AS AvgStranek,
    SUM(DobaNavstevy) AS DobaNavstevy,
    SUM(DobaNavstevy) / COUNT(Navsteva) AS AvgDobaNavstevy
FROM
    GA_data
GROUP BY 1;


-- všichni uživatelé včetně transakčních dat
SELECT
    A.Navstevnik as Navstevnik,
    A.PocetNavstev as PocetNavstev,
    A.PocetNavstivenychStranek as PocetNavstivenychStranek,
    A.AvgStranek as AvgStranek,
    A.DobaNavstevy as DobaNavstevy,
    A.AvgDobaNavstevy as AvgDobaNavstevy,
    B.PocetTransakci as PocetTransakci,
    B.Revenue as Revenue,
    B.AvgRevenue as AvgRevenue,
    B.MinRevenue as MinRevenue,
    B.MaxRevenue as MaxRevenue,
    B.KoefTransakceMesic as KoefTransakceMesic
FROM
    Uzivatele A
LEFT JOIN TransakceAGR B ON A.Navstevnik=B.Navstevnik
;

-- všichni uživatelé včetně transakčních dat a revenue po produktech
SELECT
    A.Navstevnik as Navstevnik,
    A.PocetNavstev as PocetNavstev,
    A.PocetNavstivenychStranek as PocetNavstivenychStranek,
    A.AvgStranek as AvgStranek,
    A.DobaNavstevy as DobaNavstevy,
    A.AvgDobaNavstevy as AvgDobaNavstevy,
    B.PocetTransakci as PocetTransakci,
    B.Revenue as Revenue,
    B.AvgRevenue as AvgRevenue,
    B.MinRevenue as MinRevenue,
```

```
    B.MaxRevenue as MaxRevenue,
    B.KoefTransakceMesic as KoefTransakceMesic,
    C.Obleceni as Obleceni,
    C.Doplnky as Doplnky,
    C.Elektro as Elektro,
    C.KancelarskePotreby as KancelarskePotreby,
    C.KabelkyTasky as KabelkyTasky,
    C.Brand as Brand,
    C.SkleniceHrnky as SkleniceHrnky,
    C.ZivotniStyl as ZivotniStyl,
    D.KanalDirect as KanalDirect,
    D.KanalOrganicSearch as KanalOrganicSearch,
    D.KanalPaidSearch as KanalPaidSearch,
    D.KanalSocial as KanalSocial,
    D.KanalAffiliates as KanalAffiliates,
    D.KanalDisplay as KanalDisplay,
    D.KanalOstatni as KanalOstatni,
    E.ZarizeniDesktop as ZarizeniDesktop,
    E.ZarizeniMobil as ZarizeniMobil,
    E.ZarizeniTablet as ZarizeniTablet
FROM
    Uzivatele A
LEFT JOIN TransakceAGR B ON A.Navstevnik=B.Navstevnik
LEFT JOIN ProduktyRevenue C ON A.Navstevnik=C.Navstevnik
LEFT JOIN KanalyRevenue D ON A.Navstevnik=D.Navstevnik
LEFT JOIN ZarizeniRevenue E ON A.Navstevnik=E.Navstevnik
;
```

Příloha 2: Dílčí výsledky a souhrny

Souhrn pro kategorie

Kateg..	PocetNavstev	PocetTransakci	AVG.Trans/Mesic	Median Revenue	Min. Revenue	Max. Revenue	KonverzniPomer	AVG.Time	AVG.Pages
(entrance)	236,243	4	0	267	86	577	0,00%	134	4
(not set)	221,912	699	58	83	3	66,182	0,30%	352	9
Accessories	52,369	810	68	184	3	26,862	0,91%	1,188	32
Apparel	141,562	5,194	433	82	2	477,526	1,86%	1,172	31
Bags	72,929	1,182	98	165	6	285,752	0,88%	1,230	34
Brands	77,068	827	69	79	3	22,405	0,68%	743	19
Drinkware	51,364	799	67	126	2	231,365	0,97%	1,167	32
Electronics	65,698	633	53	181	3	74,166	0,60%	1,064	29
Lifestyle	11,379	140	12	279	7	9,917	0,75%	1,288	41
Nest	409	1	0	158	158	158	0,23%	697	11
Office	54,076	1,457	121	247	2	336,688	1,53%	1,403	39

Souhrn pro zařízení a zdroj návštěvy

Kategorie P..	Zarizeni	Zdroj Navstevy	PocetNavstev	PocetTransakci	AVG.Trans/Mesic	Median Revenue	Min. Revenue	Max. Revenue	KonverzniPomer	AVG.Time	AVG.Pages
Office	desktop	(Other)	7	0	0				0,00%	1,604	43
		Affiliates	916	3	0	92	13	179	0,21%	850	17
		Direct	7,528	290	24	407	5	336,688	2,03%	1,669	45
		Display	478	18	2	390	23	3,642	1,93%	1,543	45
		Organic Search	20,463	360	30	159	5	46,510	1,02%	1,346	38
		Paid Search	1,546	49	4	256	7	2,755	1,72%	1,500	46
		Referral	10,747	632	53	278	2	72,909	3,06%	1,659	48
		Social	1,429	5	0	172	51	425	0,24%	899	24
	mobile	(Other)	5	1	0	12	12	12	10,00%	777	29
		Affiliates	51	0	0				0,00%	713	20
		Direct	2,050	32	3	166	9	16,475	0,94%	1,212	30
		Display	131	3	0	82	21	5,226	1,48%	1,098	28
		Organic Search	5,550	32	3	92	4	1,080	0,36%	1,113	29
		Paid Search	756	13	1	32	12	1,765	0,99%	1,417	36
		Referral	282	0	0				0,00%	1,036	26
		Social	341	1	0				0,20%	851	24
	tablet	Affiliates	7	0	0				0,00%	1,088	35
		Direct	349	5	0	339	22	1,824	0,88%	1,382	33
		Display	20	0	0				0,00%	894	29
		Organic Search	1,233	10	1	122	26	1,798	0,51%	1,173	33
		Paid Search	136	2	0	2,238	102	4,374	0,83%	1,445	40
		Referral	53	0	0				0,00%	937	29
		Social	54	1	0	209	209	209	1,08%	2,020	82

Souhrn pro zařízení a zdroj návštěvy

Kategorie P..	Zarizeni	Zdroj Navstevy	PocetNavstev	PocetTransakci	AVG.Trans/Mesic	Median Revenue	Min. Revenue	Max. Revenue	KonverzniPomer	AVG.Time	AVG.Pages
Lifestyle	desktop	Display	100	0	0				0,00%	1,028	40
		Organic Search	3,953	32	3	417	13	3,888	0,50%	1,339	41
		Paid Search	252	0	0				0,00%	1,386	52
		Referral	3,182	74	6	224	7	9,917	1,39%	1,294	45
		Social	263	0	0				0,00%	921	32
	mobile	Affiliates	3	0	0				0,00%	976	50
		Direct	268	3	0	42	7	95	0,65%	1,498	40
		Display	7	0	0				0,00%	1,021	34
		Organic Search	811	8	1	47	9	789	0,63%	1,039	32
		Paid Search	119	1	0	237	237	237	0,53%	1,426	37
		Referral	31	0	0				0,00%	1,442	38
		Social	32	0	0				0,00%	1,160	33
	tablet	Affiliates	2	0	0				0,00%	1,347	29
		Direct	46	0	0				0,00%	1,692	55
		Display	1	0	0				0,00%	284	8
		Organic Search	156	0	0				0,00%	1,090	35
		Paid Search	18	0	0				0,00%	2,563	78
		Referral	5	0	0				0,00%	415	15
		Social	7	1	0	204	204	204	7,14%	7,765	307

Souhrn pro zařízení a zdroj návštěvy

Kategorie P..	Zarizeni	Zdroj Navstevy	PocetNavstev	PocetTransakci	AVG.Trans/Mesic	Median Revenue	Min. Revenue	Max. Revenue	KonverzniPomer	AVG.Time	AVG.Pages
Electronics	desktop	Direct	9,427	119	10	247	3	74,166	0,74%	1,230	32
		Display	524	4	0	323	88	4,683	0,44%	1,237	33
		Organic Search	23,053	168	14	241	6	35,989	0,46%	1,085	30
		Paid Search	1,748	17	1	237	14	28,753	0,57%	1,140	36
		Referral	14,875	268	22	177	4	24,627	1,11%	1,082	32
		Social	2,107	8	1	75	25	1,209	0,24%	897	24
	mobile	(Other)	6	0	0				0,00%	556	20
		Affiliates	73	0	0				0,00%	879	25
		Direct	2,715	6	1	40	19	82	0,14%	901	23
		Display	150	2	0	377	106	648	0,73%	1,175	30
		Organic Search	6,293	24	2	137	12	1,020	0,26%	878	22
		Paid Search	852	5	0	140	10	558	0,36%	1,057	28
		Referral	304	0	0				0,00%	847	24
		Social	493	1	0	54	54	54	0,13%	712	21
	tablet	Affiliates	9	0	0				0,00%	1,817	38
		Direct	399	1	0	278	278	278	0,16%	1,073	28
		Display	22	0	0				0,00%	1,256	21
		Organic Search	1,249	8	1	34	9	186	0,43%	971	27
		Paid Search	186	1	0	129	129	129	0,35%	909	26
		Referral	54	0	0				0,00%	1,118	28
		Social	58	1	0	809	809	809	1,14%	1,630	53

Souhrn pro zařízení a zdroj návštěvy

Kategorie P..	Zarizeni	Zdroj Navstevy	PocetNavstev	PocetTransakci	AVG.Trans/Mesic	Median Revenue	Min. Revenue	Max. Revenue	KonverzniPomer	AVG.Time	AVG.Pages
Drinkware	desktop	Affiliates	909	1	0	2,540	2,540	2,540	0,07%	890	17
		Direct	7,694	116	10	249	7	12,150	0,91%	1,263	34
		Display	470	16	1	316	70	231,365	2,02%	1,181	35
		Organic Search	18,757	195	16	111	2	12,280	0,67%	1,115	30
		Paid Search	1,409	21	2	61	12	6,810	0,93%	1,167	36
		Referral	11,156	373	31	153	4	61,339	1,94%	1,387	41
		Social	1,251	3	0	282	67	2,232	0,17%	867	24
	mobile	(Other)	2	0	0				0,00%	607	5
		Affiliates	40	0	0				0,00%	608	20
		Direct	2,055	14	1	87	12	7,135	0,45%	958	23
		Display	92	1	0	71	71	71	0,66%	1,556	35
		Organic Search	5,007	37	3	33	6	731	0,47%	979	24
		Paid Search	674	7	1	71	20	1,854	0,63%	1,188	29
		Referral	174	1	0	543	543	543	0,41%	866	24
		Social	346	0	0				0,00%	685	19
	tablet	Affiliates	6	0	0				0,00%	1,129	34
		Direct	302	2	0	20	19	22	0,41%	1,258	33
		Display	13	0	0				0,00%	859	16
		Organic Search	870	9	1	38	15	594	0,67%	1,045	30
		Paid Search	118	2	0	45	20	70	0,95%	1,324	37
		Referral	38	1	0	28	28	28	1,56%	1,230	28
		Social	34	0	0				0,00%	761	35

Souhrn pro zařízení a zdroj návštěvy

Kategorie P..	Zarizeni	Zdroj Navstevy	PocetNavstev	PocetTransakci	AVG.Trans/Mesic	Median Revenue	Min. Revenue	Max. Revenue	KonverzniPomer	AVG.Time	AVG.Pages
Brands	desktop	Affiliates	599	0	0				0,00%	940	17
		Direct	5,668	95	8	100	3	17,013	1,07%	1,028	25
		Display	320	7	1	289	37	4,970	1,36%	1,035	29
		Organic Search	26,989	282	23	64	3	22,405	0,65%	787	20
		Paid Search	1,307	25	2	60	13	5,523	1,27%	882	26
		Referral	8,813	239	20	172	13	16,350	1,80%	992	27
		Social	5,058	9	1	34	12	228	0,12%	482	13
	mobile	(Other)	4	0	0				0,00%	240	12
		Affiliates	91	0	0				0,00%	615	22
		Direct	3,939	38	3	31	6	375	0,60%	679	17
		Display	122	0	0				0,00%	882	23
		Organic Search	16,243	103	9	30	6	1,601	0,40%	573	15
		Paid Search	842	7	1	92	18	540	0,53%	879	22
		Referral	390	1	0	27	27	27	0,16%	737	21
		Social	2,235	1	0	13	13	13	0,03%	488	14
	tablet	Affiliates	16	0	0				0,00%	777	26
		Direct	593	4	0	38	17	138	0,45%	730	19
		Display	17	0	0				0,00%	559	13
		Organic Search	3,355	16	1	72	14	2,208	0,32%	578	17
		Paid Search	178	0	0				0,00%	757	22
		Referral	54	0	0				0,00%	766	21
		Social	330	0	0				0,00%	351	13

Souhrn pro zařízení a zdroj návštěvy

Kategorie P..	Zarizeni	Zdroj Navstevy	PocetNavstev	PocetTransakci	AVG.Trans/Mesic	Median Revenue	Min. Revenue	Max. Revenue	KonverzniPomer	AVG.Time	AVG.Pages
Bags	desktop	Display	748	11	1	629	21	14,469	0,77%	1,102	32
		Organic Search	26,565	244	20	183	7	31,465	0,51%	1,200	33
		Paid Search	1,922	29	2	177	10	9,806	0,79%	1,324	42
		Referral	18,153	636	53	158	6	40,424	1,71%	1,405	41
		Social	1,699	7	1	170	26	394	0,27%	897	24
	mobile	(Other)	3	0	0				0,00%	696	34
		Affiliates	33	0	0				0,00%	659	21
		Direct	2,350	12	1	107	22	6,365	0,31%	1,012	25
		Display	164	0	0				0,00%	830	21
		Organic Search	6,823	35	3	104	15	1,509	0,31%	949	25
		Paid Search	807	6	1	90	16	105	0,43%	1,182	29
		Referral	227	1	0	475	475	475	0,28%	1,033	24
		Social	334	0	0				0,00%	767	22
	tablet	(Other)	2	0	0				0,00%	1,320	33
		Affiliates	8	0	0				0,00%	2,138	46
		Direct	339	2	0	82	28	137	0,37%	1,096	30
		Display	14	0	0				0,00%	1,863	30
		Organic Search	1,184	11	1	140	18	1,286	0,58%	1,107	30
		Paid Search	141	1	0	66	66	66	0,40%	1,310	34
		Referral	46	3	0	107	96	389	3,66%	1,174	31
		Social	50	0	0				0,00%	427	16

Souhrn pro zařízení a zdroj návštěvy

Kategorie P..	Zarizeni	Zdroj Navstevy	PocetNavstev	PocetTransakci	AVG.Trans/Mesic	Median Revenue	Min. Revenue	Max. Revenue	KonverzniPomer	AVG.Time	AVG.Pages
Apparel	desktop	Affiliates	1,600	2	0	30	30	31	0,07%	955	17
		Direct	16,532	722	60	82	3	405,469	2,21%	1,327	33
		Display	1,199	57	5	171	11	477,526	2,28%	1,351	36
		Organic Search	49,974	1,256	105	78	3	46,636	1,30%	1,107	29
		Paid Search	3,889	177	15	82	10	14,975	2,19%	1,207	36
		Referral	29,322	2,500	208	94	2	44,144	3,62%	1,621	45
		Social	3,477	64	5	42	7	3,411	1,17%	850	22
	mobile	(Other)	9	0	0				0,00%	593	12
		Affiliates	136	1	0	14	14	14	0,47%	707	19
		Direct	6,207	72	6	63	5	24,409	0,69%	844	21
		Display	287	2	0	57	25	90	0,37%	1,007	26
		Organic Search	20,482	228	19	32	2	1,858	0,63%	800	21
		Paid Search	2,305	43	4	59	12	3,192	1,06%	976	24
		Referral	621	3	0	76	50	77	0,29%	925	24
		Social	1,308	2	0	233	50	416	0,10%	679	18
	tablet	(Other)	1	0	0				0,00%	566	14
		Affiliates	20	0	0				0,00%	740	17
		Direct	787	12	1	40	12	254	0,89%	1,122	29
		Display	20	0	0				0,00%	805	26
		Organic Search	3,088	43	4	84	16	1,662	0,81%	997	27
		Paid Search	377	8	1	69	24	1,190	1,15%	987	29
		Referral	101	1	0	23	23	23	0,76%	666	16
		Social	126	1	0	26	26	26	0,48%	1,045	37

Shluk cluster_2

Navstevnik	Pocet Navstev	Pocet Transakci	Avg. Avg Doba N..	Avg. Avg Stranek	Avg. Avg Revenue	Revenue	Obleceni	Kancelarske Pot..	Kabelky Tasky	Elektro	Doplnky	Sklenice Hrnky
992792274..	22	0	1,752	23	0	0	0	0	0	0	0	0
992976518..	74	0	3,729	92	0	0	0	0	0	0	0	0
993668261..	61	6	7,814	83	162	974	144	0	116	335	0	0
993782157..	14	0	2,435	27	0	0	0	0	0	0	0	0
993931629..	27	0	1,666	17	0	0	0	0	0	0	0	0
994612906..	24	2	5,663	103	4,437	8,873	0	6,702	0	0	0	0
995018230..	14	0	3,195	31	0	0	0	0	0	0	0	0
995261617..	97	3	925	16	2,168	6,504	6,504	0	0	0	0	0
995629614..	23	1	1,892	28	4,745	4,745	4,745	0	0	0	0	0
996063716..	18	0	3,140	96	0	0	0	0	0	0	0	0
996383485..	27	0	1,430	46	0	0	0	0	0	0	0	0
996742146..	43	1	1,647	41	503	503	0	0	0	0	0	0
996789649..	14	0	2,900	38	0	0	0	0	0	0	0	0
996803515..	41	1	1,750	34	240	240	0	0	0	0	240	0
997037802..	38	3	924	25	121	363	0	193	22	0	0	148
997274198..	22	0	2,035	89	0	0	0	0	0	0	0	0
997358958..	59	0	563	33	0	0	0	0	0	0	0	0
997423225..	68	3	1,884	42	7,092	21,277	0	21,277	0	0	0	0
997435191..	18	1	2,035	34	173	173	173	0	0	0	0	0
997570217..	38	1	940	34	85	85	0	0	0	0	0	85
997712382..	30	2	1,935	64	72	144	144	0	0	0	0	0
998156222..	22	1	5,381	90	215	215	0	0	215	0	0	0
998157580..	12	0	3,166	37	0	0	0	0	0	0	0	0
998159354..	20	0	1,878	22	0	0	0	0	0	0	0	0
998298577..	40	0	1,122	25	0	0	0	0	0	0	0	0
999017446..	19	0	3,859	76	0	0	0	0	0	0	0	0
999018361..	55	2	904	11	226	452	0	0	399	0	0	0
999476707..	82	1	1,503	33	842	842	0	842	0	0	0	0
999740924..	24	1	1,423	65	202	202	0	202	0	0	0	0
Grand Total	125,943	2,672	2,064	43	843	5,648,797	1,764,963	1,233,227	783,137	550,119	488,889	349,886

Shluk cluster_2

Navstevnik	Brand	Zivotni Styl	Kanal Direct	Kanal Referral	Kanal Organic S..	Kanal Paid Search	Kanal Social	Kanal Ostatni	Kanal Display	Kanal Affiliates	Zarizeni Desktop	Zarizeni Mobil
992792274..	0	0	0	0	0	0	0	0	0	0	0	0
992976518..	0	0	0	0	0	0	0	0	0	0	0	0
993668261..	379	0	974	0	0	0	0	0	0	0	974	0
993782157..	0	0	0	0	0	0	0	0	0	0	0	0
993931629..	0	0	0	0	0	0	0	0	0	0	0	0
994612906..	0	0	0	0	6,702	2,171	0	0	0	0	8,873	0
995018230..	0	0	0	0	0	0	0	0	0	0	0	0
995261617..	0	0	6,504	0	0	0	0	0	0	0	6,504	0
995629614..	0	0	0	0	4,745	0	0	0	0	0	4,745	0
996063716..	0	0	0	0	0	0	0	0	0	0	0	0
996383485..	0	0	0	0	0	0	0	0	0	0	0	0
996742146..	503	0	0	0	503	0	0	0	0	0	0	503
996789649..	0	0	0	0	0	0	0	0	0	0	0	0
996803515..	0	0	0	0	240	0	0	0	0	0	240	0
997037802..	0	0	0	363	0	0	0	0	0	0	363	0
997274198..	0	0	0	0	0	0	0	0	0	0	0	0
997358958..	0	0	0	0	0	0	0	0	0	0	0	0
997423225..	0	0	21,277	0	0	0	0	0	0	0	21,277	0
997435191..	0	0	0	0	173	0	0	0	0	0	173	0
997570217..	0	0	0	85	0	0	0	0	0	0	85	0
997712382..	0	0	0	0	144	0	0	0	0	0	144	0
998156222..	0	0	0	215	0	0	0	0	0	0	215	0
998157580..	0	0	0	0	0	0	0	0	0	0	0	0
998159354..	0	0	0	0	0	0	0	0	0	0	0	0
998298577..	0	0	0	0	0	0	0	0	0	0	0	0
999017446..	0	0	0	0	0	0	0	0	0	0	0	0
999018361..	53	0	0	0	452	0	0	0	0	0	452	0
999476707..	0	0	0	0	842	0	0	0	0	0	842	0
999740924..	0	0	0	202	0	0	0	0	0	0	202	0
Grand Total	188,707	69	2,378,603	1,785,648	1,249,176	205,084	0	0	0	0	5,437,563	192,224

Shluk cluster_1

Navstevnik	Pocet Navstev	Pocet Transakci	Avg. Avg Doba Navstevy	Avg. Avg Stranek	Avg. Avg Revenue	Revenue	Obleceni	Kancelarske Potreby	Kabelky Tasky	Elektro	Doplnky	Sklenice Hrnky	B
082483972..	866	0	2,962	26	0	0	0	0	0	0	0	0	
195745897..	864	18	1,641	33	77,750	1,399,494	857,445	2,508	300,221	4,683	0	232,083	
Grand Total	1,730	18	2,301	30	38,875	1,399,494	857,445	2,508	300,221	4,683	0	232,083	

Shluk cluster_1

Navstevnik	Brand	Zivotni Styl	Kanal Direct	Kanal Referral	Kanal Organic Search	Kanal Paid Search	Kanal Social	Kanal Ostatni	Kanal Display	Kanal Affiliates	Zarizeni Desktop	Zarizeni Mobil
082483972..	0	0	0	0	0	0	0	0	0	0	0	0
195745897..	0	0	285,752	0	0	0	0	0		0	1,399,494	0
Grand Total	0	0	285,752	0	0	0	0	0	0	0	1,399,494	0

Shluk cluster_0

Navstevnik	Pocet Navstev	Pocet Transakci	Avg. Avg Doba N..	Avg. Avg Stranek	Avg. Avg Revenue	Revenue	Obleceni	Kancelarske Pot..	Kabelky Tasky	Elektro	Doplnky	Sklenice Hrnky
999906095..	1	0	0	1	0	0	0	0	0	0	0	0
999909460..	4	0	485	10	0	0	0	0	0	0	0	0
999913694..	1	0	29	3	0	0	0	0	0	0	0	0
999935061..	1	0	0	1	0	0	0	0	0	0	0	0
999935558..	1	0	0	1	0	0	0	0	0	0	0	0
999943091..	6	0	902	24	0	0	0	0	0	0	0	0
999943980..	1	0	4	2	0	0	0	0	0	0	0	0
999944083..	1	0	21	2	0	0	0	0	0	0	0	0
999947022..	1	0	82	8	0	0	0	0	0	0	0	0
999950970..	1	0	0	1	0	0	0	0	0	0	0	0
999952005..	14	0	175	9	0	0	0	0	0	0	0	0
999954564..	1	0	2,350	3	0	0	0	0	0	0	0	0
999959186..	1	0	18	2	0	0	0	0	0	0	0	0
999963269..	1	0	0	1	0	0	0	0	0	0	0	0
999967908..	1	0	0	1	0	0	0	0	0	0	0	0
999973422..	1	0	0	1	0	0	0	0	0	0	0	0
999973962..	1	0	0	1	0	0	0	0	0	0	0	0
999976128..	4	0	227	8	0	0	0	0	0	0	0	0
999977307..	1	0	0	1	0	0	0	0	0	0	0	0
999977483..	1	0	0	1	0	0	0	0	0	0	0	0
999979980..	4	0	42	5	0	0	0	0	0	0	0	0
999980122..	12	0	608	23	0	0	0	0	0	0	0	0
999982406..	1	0	0	1	0	0	0	0	0	0	0	0
999982563..	6	0	592	9	0	0	0	0	0	0	0	0
999988742..	5	0	61	4	0	0	0	0	0	0	0	0
999990672..	2	0	515	4	0	0	0	0	0	0	0	0
999996318..	1	0	0	1	0	0	0	0	0	0	0	0
999997826..	1	0	0	1	0	0	0	0	0	0	0	0
999998643..	1	0	55	2	0	0	0	0	0	0	0	0
Grand Total	1,270,226	9,056	191	6	9	3,511,513	1,253,391	822,743	355,181	165,211	247,858	204,492

Shluk cluster_0

Navstevnik	Brand	Zivotni Styl	Kanal Direct	Kanal Referral	Kanal Organic S..	Kanal Paid Search	Kanal Social	Kanal Ostatni	Kanal Display	Kanal Affiliates	Zarizeni Desktop	Zarizeni Mobil	Z
999906095..	0	0	0	0	0	0	0	0	0	0	0	0	
999909460..	0	0	0	0	0	0	0	0	0	0	0	0	
999913694..	0	0	0	0	0	0	0	0	0	0	0	0	
999935061..	0	0	0	0	0	0	0	0	0	0	0	0	
999935558..	0	0	0	0	0	0	0	0	0	0	0	0	
999943091..	0	0	0	0	0	0	0	0	0	0	0	0	
999943980..	0	0	0	0	0	0	0	0	0	0	0	0	
999944083..	0	0	0	0	0	0	0	0	0	0	0	0	
999947022..	0	0	0	0	0	0	0	0	0	0	0	0	
999950970..	0	0	0	0	0	0	0	0	0	0	0	0	
999952005..	0	0	0	0	0	0	0	0	0	0	0	0	
999954564..	0	0	0	0	0	0	0	0	0	0	0	0	
999959186..	0	0	0	0	0	0	0	0	0	0	0	0	
999963269..	0	0	0	0	0	0	0	0	0	0	0	0	
999967908..	0	0	0	0	0	0	0	0	0	0	0	0	
999973422..	0	0	0	0	0	0	0	0	0	0	0	0	
999973962..	0	0	0	0	0	0	0	0	0	0	0	0	
999976128..	0	0	0	0	0	0	0	0	0	0	0	0	
999977307..	0	0	0	0	0	0	0	0	0	0	0	0	
999977483..	0	0	0	0	0	0	0	0	0	0	0	0	
999979980..	0	0	0	0	0	0	0	0	0	0	0	0	
999980122..	0	0	0	0	0	0	0	0	0	0	0	0	
999982406..	0	0	0	0	0	0	0	0	0	0	0	0	
999982563..	0	0	0	0	0	0	0	0	0	0	0	0	
999988742..	0	0	0	0	0	0	0	0	0	0	0	0	
999990672..	0	0	0	0	0	0	0	0	0	0	0	0	
999996318..	0	0	0	0	0	0	0	0	0	0	0	0	
999997826..	0	0	0	0	0	0	0	0	0	0	0	0	
999998643..	0	0	0	0	0	0	0	0	0	0	0	0	
Grand Total	90,411	477	1,173,534	1,488,964	703,649	71,910	0	0	1,177	14	3,388,340	101,037	